

MECHANICAL ENGINEERS' HANDBOOK

Volume
Manufacturing and
Management » **3**

MYER KUTZ EDITOR

FOURTH EDITION

WILEY

Mechanical Engineers' Handbook

Mechanical Engineers' Handbook
Fourth Edition

Manufacturing and Management

Edited by
Myer Kutz

WILEY

Cover image: © denisovd / Thinkstock

Cover design: Wiley

This book is printed on acid-free paper.

Copyright © 2015

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with the respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor the author shall be liable for damages arising herefrom.

For general information about our other products and services, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Library of Congress Cataloging-in-Publication Data

Mechanical engineers handbook : manufacturing and management / edited by Myer Kutz. – Fourth edition.

1 online resource.

Includes index.

Description based on print version record and CIP data provided by publisher; resource not viewed.

ISBN 978-1-118-93082-3 (ePub) – ISBN 978-1-118-93081-6 (Adobe PDF) –

ISBN 978-1-118-11899-3 (4-volume set) – ISBN 978-1-118-11284-7 (cloth : volume 3 : acid-free paper)

1. Mechanical engineering—Handbooks, manuals, etc. I. Kutz, Myer, editor of compilation.

TJ151

621-dc23

2014005952

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Alan and Nancy, now and forever

Contents

Preface	ix
Vision for the Fourth Edition	xi
Contributors	xiii

PART 1 MANUFACTURING

1

1. Organization, Management, and Improvement of Manufacturing Systems	3
<i>Keith M. Gardiner</i>	
2. Environmentally Benign Manufacturing	29
<i>William E. Biles</i>	
3. Production Planning	53
<i>Bhaba R. Sarker, Dennis B. Webster, and Thomas G. Ray</i>	
4. Production Processes and Equipment	115
<i>Magd E. Zohdi, William E. Biles, and Dennis B. Webster</i>	
5. Manufacturing Systems Evaluation	183
<i>Walter W. Olson</i>	
6. Metal Forming, Shaping, and Casting	195
<i>Magd E. Zohdi and William E. Biles</i>	
7. Coatings and Surface Engineering: Physical Vapor Deposition	235
<i>Allan Matthews and Suzanne L. Rohde</i>	
8. Mechanical Fasteners	255
<i>Murray J. Roblin and Updated by Anthony Luscher</i>	
9. Seal Technology	283
<i>Bruce M. Steinetz</i>	
10. Statistical Quality Control	325
<i>Magd E. Zohdi</i>	
11. Computer-Integrated Manufacturing	339
<i>William E. Biles and Magd E. Zohdi</i>	
12. TRIZ	361
<i>James E. McMunigal, Steven Ungvari, Michael Slocum, and Ruth E. McMunigal</i>	
13. Data Exchange Using STEP	391
<i>Martin Hardwick</i>	
14. Achieving Enterprise Goals with New Process Technology	397
<i>Steve W. Tuszynski</i>	
15. Nondestructive Inspection	441
<i>Robert L. Crane and Giles Dillingham</i>	
16. Materials Handling System Design	497
<i>Sunderesh S. Heragu and Banu Ekren</i>	

**PART 2 MANAGEMENT, FINANCE, QUALITY, LAW,
AND RESEARCH****513**

17.	Intelligent Control of Material Handling Systems	515
	<i>Kasper Hallenborg</i>	
18.	Managing People in Engineering and Technology	559
	<i>Hans J. Thamhain</i>	
19.	Engineering Economy	581
	<i>Kate D. Abel</i>	
20.	Evaluating and Selecting Technology-Based Projects	605
	<i>Hans J. Thamhain</i>	
21.	Lean Management	617
	<i>Eric H. Stapp and Cynthia M. Sabelhaus</i>	
22.	Total Quality Management for Mechanical Engineers	635
	<i>Alan Kemerling</i>	
23.	Registrations, Certifications, and Awards	667
	<i>Cynthia M. Sabelhaus and Eric H. Stapp</i>	
24.	Safety Engineering	691
	<i>Jack B. ReVelle</i>	
25.	What the Law Requires of the Engineer	749
	<i>Alvin S. Weinstein and Martin S. Chizek</i>	
26.	Patents	773
	<i>David A. Burge and Benjamin D. Burge</i>	
27.	Online Information Resources for Mechanical Engineers	805
	<i>Robert N. Schwarzwald, Jr.</i>	
28.	Sources of Mechanical Engineering Information	823
	<i>Fritz Dusold and Myer Kutz</i>	
	Index	831

Preface

The third volume of the fourth edition of the *Mechanical Engineers' Handbook* comprises two parts: Manufacturing and Management. Each part contains 12 chapters. Contributors include business owners, consultants, lawyers, librarians, and academics from all around the United States.

Part 1 opens with a chapter from the second edition on Product Design for Manufacturing and Assembly (DFM&A). The centerpiece of Part 1 includes the chapters that in earlier editions of the handbook have been called “the handbook within the handbook.”

Developed by a team at Louisiana State University and the University of Louisville, these six chapters, which have been updated, span manufacturing topics from production planning, production processes and equipment, metal forming, shaping, and casting, statistical quality control, computer-integrated manufacturing, to material handling. The chapter on classification systems remains unchanged from earlier editions; the chapter on mechanical fasteners has been revised extensively. Part 1 has three chapters entirely new to the handbook: a chapter on physical vapor deposition, one on environmentally conscious manufacturing, and one on a new approach to dealing with process technology in the context of design, tooling, manufacturing, and quality engineering. The latter chapter is indicative of how much contributors can give of themselves. Its content is the lifeblood of its author's consulting practice.

Part 2 covers a broad array of topics. The 12 chapters can be broken down into four groups. The first two chapters cover project and people management. The first of these chapters, on project management, deals with a subject that has appeared in previous editions, but the chapter is entirely new, to reflect advances in this field. The people management chapter has been revised. The following three chapters deal with fundamentals of financial management and are unchanged. The next three chapters, contributed by a team led by Jack ReVelle, treat a set of management issues, including total quality management; registrations, certifications, and awards; and safety engineering. Two chapters cover legal issues of interest to engineers, including patents. The final two chapters cover online and print information sources useful to mechanical engineers in their daily work. The chapter on online sources is a new version of the chapter that appeared originally in 1998.

Vision for the Fourth Edition

Basic engineering disciplines are not static, no matter how old and well established they are. The field of mechanical engineering is no exception. Movement within this broadly based discipline is multidimensional. Even the classic subjects, on which the discipline was founded, such as mechanics of materials and heat transfer, keep evolving. Mechanical engineers continue to be heavily involved with disciplines allied to mechanical engineering, such as industrial and manufacturing engineering, which are also constantly evolving. Advances in other major disciplines, such as electrical and electronics engineering, have significant impact on the work of mechanical engineers. New subject areas, such as neural networks, suddenly become all the rage.

In response to this exciting, dynamic atmosphere, the Mechanical Engineers' Handbook expanded dramatically, from one to four volumes for the third edition, published in November 2005. It not only incorporated updates and revisions to chapters in the second edition, published seven years earlier, but also added 24 chapters on entirely new subjects, with updates and revisions to chapters in the Handbook of Materials Selection, published in 2002, as well as to chapters in Instrumentation and Control, edited by Chester Nachtigal and published in 1990, but never updated by him.

The fourth edition retains the four-volume format, but there are several additional major changes. The second part of Volume I is now devoted entirely to topics in engineering mechanics, with the addition of five practical chapters on measurements from the Handbook of Measurement in Science and Engineering, published in 2013, and a chapter from the fifth edition of Eshbach's Handbook of Engineering Fundamentals, published in 2009. Chapters on mechanical design have been moved from Volume I to Volumes II and III. They have been augmented with four chapters (updated as needed) from Environmentally Conscious Mechanical Design, published in 2007. These chapters, together with five chapters (updated as needed, three from Environmentally Conscious Manufacturing, published in 2007, and two from Environmentally Conscious Materials Handling, published in 2009) in the beefed-up manufacturing section of Volume III, give the handbook greater and practical emphasis on the vital issue of sustainability.

Prefaces to the handbook's individual volumes provide further details on chapter additions, updates and replacements. The four volumes of the fourth edition are arranged as follows:

- Volume 1: Materials and Engineering Mechanics—27 chapters
 - Part 1. Materials—15 chapters
 - Part 2. Engineering Mechanics—12 chapters
- Volume 2: Design, Instrumentation and Controls—25 chapters
 - Part 1. Mechanical Design—14 chapters
 - Part 2. Instrumentation, Systems, Controls and MEMS —11 chapters
- Volume 3: Manufacturing and Management—28 chapters
 - Part 1. Manufacturing—16 chapters
 - Part 2. Management, Finance, Quality, Law, and Research—12 chapters
- Volume 4: Energy and Power—35 chapters
 - Part 1: Energy—16 chapters
 - Part 2: Power—19 chapters

The mechanical engineering literature is extensive and has been so for a considerable period of time. Many textbooks, reference works, and manuals as well as a substantial number of journals exist. Numerous commercial publishers and professional societies, particularly in the United States and Europe, distribute these materials. The literature grows continuously, as applied mechanical engineering research finds new ways of designing, controlling, measuring, making, and maintaining things, as well as monitoring and evaluating technologies, infrastructures, and systems.

Most professional-level mechanical engineering publications tend to be specialized, directed to the specific needs of particular groups of practitioners. Overall, however, the mechanical engineering audience is broad and multidisciplinary. Practitioners work in a variety of organizations, including institutions of higher learning, design, manufacturing, and consulting firms, as well as federal, state, and local government agencies. A rationale for a general mechanical engineering handbook is that every practitioner, researcher, and bureaucrat cannot be an expert on every topic, especially in so broad and multidisciplinary a field, and may need an authoritative professional summary of a subject with which he or she is not intimately familiar.

Starting with the first edition, published in 1986, my intention has always been that the Mechanical Engineers' Handbook stand at the intersection of textbooks, research papers, and design manuals. For example, I want the handbook to help young engineers move from the college classroom to the professional office and laboratory where they may have to deal with issues and problems in areas they have not studied extensively in school.

With this fourth edition, I have continued to produce a practical reference for the mechanical engineer who is seeking to answer a question, solve a problem, reduce a cost, or improve a system or facility. The handbook is not a research monograph. Its chapters offer design techniques, illustrate successful applications, or provide guidelines to improving performance, life expectancy, effectiveness, or usefulness of parts, assemblies, and systems. The purpose is to show readers what options are available in a particular situation and which option they might choose to solve problems at hand.

The aim of this handbook is to serve as a source of practical advice to readers. I hope that the handbook will be the first information resource a practicing engineer consults when faced with a new problem or opportunity—even before turning to other print sources, even officially sanctioned ones, or to sites on the Internet. In each chapter, the reader should feel that he or she is in the hands of an experienced consultant who is providing sensible advice that can lead to beneficial action and results.

Can a single handbook, even spread out over four volumes, cover this broad, interdisciplinary field? I have designed the Mechanical Engineers' Handbook as if it were serving as a core for an Internet-based information source. Many chapters in the handbook point readers to information sources on the Web dealing with the subjects addressed. Furthermore, where appropriate, enough analytical techniques and data are provided to allow the reader to employ a preliminary approach to solving problems.

The contributors have written, to the extent their backgrounds and capabilities make possible, in a style that reflects practical discussion informed by real-world experience. I would like readers to feel that they are in the presence of experienced teachers and consultants who know about the multiplicity of technical issues that impinge on any topic within mechanical engineering. At the same time, the level is such that students and recent graduates can find the handbook as accessible as experienced engineers.

Contributors

Kate D. Abel
Stevens Institute of Technology
Hoboken, New Jersey

William E. Biles
University of Louisville
Louisville, Kentucky

Benjamin D. Burge
Intel Americas, Inc.
Chantilly, Virginia

David A. Burge
David A. Burge Company
Cleveland, Ohio

Martin S. Chizek
Weinstein Associates International
Delray Beach, Florida

Robert L. Crane
Air Force Research Laboratory
Wright Patterson Air Force Base
Dayton, Ohio

Giles Dillingham
Brighton Technologies Group
Cincinnati, Ohio

Fritz Dusold
Mid-Manhattan Library Science
and Business Department (Retired)
New York, New York

Banu Ekren
University of Louisville
Louisville, Kentucky

Keith M. Gardiner
Lehigh University
Bethlehem, Pennsylvania

Kasper Hallenborg
University of Southern Denmark
Odense, Denmark

Martin Hardwick
Rensselaer Polytechnic Institute &
STEP Tools, Inc.
Troy, New York

Sunderesh S. Heragu
University of Louisville
Louisville, Kentucky

Jeremy S. Knopp
Air Force Research Laboratory
Wright Patterson Air Force Base
Dayton, Ohio

Alan Kemerling
Ethicon, Inc.

Myer Kutz
Myer Kutz Associates, Inc.
Delmar, New York

Anthony Luscher
Ohio State University
Columbus, Ohio

Allan Matthews
Sheffield University
Sheffield, United Kingdom

James E. McMunigal
MCM Associates
Long Beach, California

Ruth E. McMunigal
MCM Associates
Long Beach, California

Walter W. Olson
University of Toledo
Toledo, Ohio

xiv Contributors

Thomas G. Ray
Louisiana State University
Baton Rouge, Louisiana

Jack B. ReVelle
Revelle Solutions, LLC
Santa Ana, California

Murray J. Roblin
California State Polytechnic University
Pomona, California

Suzanne L. Rohde
Infinidium, LLC
Steamboat Spring, Colorado

Cynthia M. Sabelhaus
Raytheon Missile Systems Company
Tucson, Arizona

Bhaba R. Sarker
Louisiana State University
Baton Rouge, Louisiana

Robert N. Schwarzwald, Jr.
Stanford University
Stanford, California

Michael Slocum
Breakthrough Management Group
Longmont, Colorado

Bruce M. Steinetz
NASA Glenn Research Center at Lewis Field
Cleveland, Ohio

Eric H. Stapp
Raytheon Missile Systems Company
Tucson, Arizona

Hans J. Thamhain
Bentley University
Waltham, Massachusetts

Steve W. Tuszynski
Algoryx, Inc.
Los Angeles, California

Steven Ungvari
Strategic Product Innovations, Inc.
Columbus, Ohio

Dennis B. Webster
Louisiana State University
Baton Rouge, Louisiana

Alvin S. Weinstein
Weinstein Associates International
Delray Beach, Florida

Magd E. Zohdi
Louisiana State University
Baton Rouge, Louisiana

PART 1

MANUFACTURING

CHAPTER 1

ORGANIZATION, MANAGEMENT, AND IMPROVEMENT OF MANUFACTURING SYSTEMS

Keith M. Gardiner
Lehigh University
Bethlehem, Pennsylvania

1	INTRODUCTION: WHAT IS THIS CHAPTER ABOUT?	3	8	WORKFORCE CONSIDERATIONS: SOCIAL ENGINEERING, THE DIFFICULT PART	15
2	NATURE OF MANUFACTURING SYSTEM: ARENA FOR OUR IMPROVEMENT	4	9	ENVIRONMENTAL CONSCIOUSNESS: MANUFACTURING EMBEDDED IN SOCIETY	17
3	EVOLUTION OF LEADERSHIP AND MANAGEMENT: HANDICAP OF HIERARCHIES	6	9.1	Sustainability	18
4	ORGANIZATIONAL BEHAVIORS, CHANGE, AND SPORTS: FRUITLESS QUEST FOR STABILITY	8	9.2	Principles for Environmentally Conscious Design	19
5	SYSTEM OF MEASUREMENT AND ORGANIZATION: STIMULATING CHANGE	10	10	IMPLEMENTATION: CONSIDERATIONS AND EXAMPLES FOR COMPANIES OF ALL SIZES	20
6	COMPONENTS OF MANUFACTURING SYSTEM: SIMPLIFIED WAY OF LOOKING AT SYSTEM	12	10.1	Vertical Integration	20
7	IMPROVEMENT, PROBLEM SOLVING, AND SYSTEMS DESIGN: ALL-EMBRACING RECYCLING, REPEATING, SPIRALING CREATIVE PROCESS	13	10.2	Real-World Examples	20
			10.3	Education Programs	22
			10.4	Measuring Results	23
			11	A LOOK TO THE FUTURE	24
				REFERENCES	26

1 INTRODUCTION: WHAT IS THIS CHAPTER ABOUT?

There are many books, pricey consultants, guides, expensive courses, and magazine articles telling us how to improve. Improvers tell us how to do everything from diet, exercise, staying healthy, relaxing, sleeping, investing, fixing our homes, and growing vegetables to bringing up our children—there are recommended fixes available for every human condition! This trend is

4 Organization, Management, and Improvement of Manufacturing Systems

nowhere more prevalent than in business and industry and most especially in manufacturing. The challenge for this chapter is to deliver meaningful content that, if applied diligently, will enable readers to improve their manufacturing systems.

We must go beyond the acronyms and buzzwords, and here there are strong parallels with self-improvement. To be successful, self-improvement and a diet or exercise regimen first requires admission, recognition, and consciousness of the necessity for improvement. The next step required is to realize that improvement is possible; then there must be a willingness and eager enthusiasm to meet the challenges and commence the task or tasks; this can be very difficult. It is too easy for managers or erstwhile change agents to place placards by the coffee and soda machines and in the cafeteria with messages like “Learn today and be here tomorrow.” Inspirational posters, T-shirts, and baseball caps with logos and slogans are often made available as promotional incentives. This is ignorant folly and can rapidly turn any improvement project into a cliché and workplace joke.

A leading slogan (maybe some slogans are unavoidable) is continuous improvement. Here the models from sports or the arts are appropriate. Athletes and musicians practice, learn, and train, almost as a way of life. Similar approaches and habits must be introduced to the manufacturing regimen. Here, management must lead by example and act as coaches while at the same time accepting that they also must be engaged in continuing endeavors to improve. Commitment and the enthusiasm of management, accompanied by visible participation, are essential. In fact, no improvement initiative should be launched without a prior thoroughgoing and preferably independent objective analysis to assess the morale of the whole operation or enterprise. Incorrect assumptions by leadership will result in poor planning, possibly inappropriate emphasis, and ineffective implementation. As a consequence there could be negative effects on workplace morale, and the initiative could be destined for failure.

Beyond this it is wise to recognize that any initiative will inevitably have a life cycle.¹ Thus, planning and implementation must be very careful and deliberate. Initiatives of this nature should not be considered as once and done. There must be long-range plans for continuation, revitalization, and refreshment. To be successful, the improvement initiative(s) must become embedded into the culture and practices of the enterprise. It must become a habit, and resources must be allocated to support successful implementation and on-going maintenance.

Improvement can be an abstract notion, but any improvement must be accompanied by a thorough analysis and understanding of exactly what is to be improved. An athlete has many performance metrics, such as resting pulse, heart and lung capacities, treadmill and weight performances, times for standard tests, and ultimately, of course, competitive results. Practice and training regimens are developed to focus on areas of weakness and to develop greater capabilities in zones of opportunity. Time is spent in counseling, measuring, and planning with development of very specific exercises on a continuing basis. It is rare to discover this kind of detailed attention being paid to the improvement of individuals, teams, or their performance in manufacturing enterprises. Nevertheless this is an essential concomitant to any improvement regions.

2 NATURE OF MANUFACTURING SYSTEM: ARENA FOR OUR IMPROVEMENT

Systems for manufacture, or production, have evolved appreciably in the last 4000 or so years. The achievements of the Egyptians, Persians, Greeks, Romans, and others must not be ignored. They were able to leave us countless superbly manufactured artifacts and equip their military as efficient conquerors. It is interesting and worthwhile to define the production or manufacturing system in this context. Our system can be viewed as “a system whereby resources (including materials and energy) are transformed to produce goods (and/or services) with generation of wealth.”² Our current systems, recent developments, and, particularly, prejudices can be best appreciated and understood by taking a brief glance back in time to review the nature, management, and characteristics of some of these early production systems.

Most early systems were directed and under the control of local rulers. In many locations these pharaohs, princes, chieftains, or tribal leaders levied taxes for defense and other purposes of state and also to support their military, social, and manufacturing systems. In Europe, after the fall of the Roman Empire, a distributed regional, state, or manorial system arose that was hierarchical. The local earls, dukes, princes, or lords of the manor owed allegiance and paid taxes to the next levels, the church, and/or threatening despots. This manorial system relied on a tiered dependent and subservient vassal or peasant society. The manor, district, or local manager (or seigneur) gave protection and loans of land to the vassals proportional to perceptions of their contribution to the unit.³ Products required for daily living, agriculture, clothing, food, meat, and fuel were produced as ordered, assuming weather and other conditions were satisfactory.

Major large-scale projects to meet architectural, marine, defense, societal, and funereal purposes (harbors, fortifications, aqueducts, and memorial structures) involved substantial mobilization of resources and possibly the use of slaves captured in wars. Smaller artifacts were made by single artisans or by small groups working collectively; agricultural production was also relatively small scale and primarily for local markets. In these early days the idea of an enterprise was synonymous with the city or city-state itself. When the armies needed equipment, swords, and armor, orders were posted and groups of artisans worked to fill them. Organization during these periods was hierarchical and devolved around the state and a ruling class. Religion also played a major role in structuring the lives of the populace.

The artisan groups organized themselves into guilds establishing standards for their craft, together with differentiation, fellowship, and support for those admitted to full membership. There was training for apprentices and aid for widows and orphans when a member died. Guilds participated actively in the religious life of the community, built almshouses, and did charitable works.⁴ It can be surmised that guild leaders of the miners in Saxony, for example, would have the power, experience, and qualifications to negotiate working conditions with the lord of the manor or leader of the principality and mine owner. The guild would also claim some share in the revenues of the mining and metal winning operations. Mining and manufacturing operations in Saxony were described extensively in *De Re Metallica*, a notable text by Agricola in 1556 translated into English by the Hoovers.⁵

The guild workplaces, mines, smelters, waterwheel-powered forges, hammers (described by Agricola), grist mills, and the like were the early factories. The existence of a water-powered paper mill in England is recorded as early as 1494. The printing operations of Gutenberg in what was to become Germany and of Caxton in England in 1454 and 1474, respectively, were small factories. Early armorers must have worked in groups supported by cupolas, furnaces, hearths, and power systems. A most renowned early factory was the Arsenale (arsenal) in Venice. This was a dockyard operated by the city-state that opened around the eighth century, with major new structures (Arsenale Nuovo) started in 1320. At its height in the sixteenth century, the arsenal was capable of producing one ship per day using an assembly line with mass production methods, prefabrication of standardized parts, division of labor, and specialization.⁶ Power sources during these periods were limited to levers, winches, and cranes driven by human or animal power, wind, or water. To a large extent these systems were reasonably sustainable but were vulnerable to unpredictable social, climatic, or other disasters.

During the period marked as the Industrial Revolution, available power densities increased markedly. Improvements in engineering and materials increased the efficiency and size of waterwheels and their associated transmission systems. There is a tendency, certainly in the United Kingdom and United States, to mark the improvement of the steam engine by Boulton and Watt and the discussions of the Lunar Society as the inception of the Industrial Revolution.⁷ In fact, effective production systems were already extant and evolving as the result of global influences. The scale and scope increased as a result of this major change in available power density. Factories grew up around sources of power, materials, and potential employees.

3 EVOLUTION OF LEADERSHIP AND MANAGEMENT: HANDICAP OF HIERARCHIES

History has given us effective models for the organization of our manufacturing systems. The notion of the paid worker as a vassal has tended to predominate, notwithstanding the wise thoughts of Adam Smith, predating W. Edwards Deming* by almost 200 years.⁸ He expressed the need for the workforce to be positively integrated as a factor engaged in the furtherance of the objectives of the manufacturing system as follows:

But what improves the circumstances of the greater part can never be regarded as an inconvenience to the whole. No society can surely be flourishing and happy, of which the far greater part of the members are poor and miserable. It is but equity, besides, that they who feed, clothe, and lodge the whole body of the people, should have such a share of the produce of their own labor as to be themselves tolerably well fed, clothed, and lodged. The liberal reward of labor, as it encourages the propagation, so it increases the industry of the common people. The wages of labor are the encouragement of industry, which, like every other human quality, improves in proportion to the encouragement it receives. A plentiful subsistence increases the bodily strength of the laborer, and the comfortable hope of bettering his condition, and of ending his days perhaps in ease and plenty, animates him to exert that strength to the utmost. Where wages are high, accordingly, we shall always find the workmen more active, diligent, and expeditious than where they are low.

It is clear that an understanding of physical, economic, social, organizational, and behavioral processes are an important aspect for the whole manufacturing or production enterprise.

And, of course, if we combed the words of Machiavelli in *The Prince* or Sun Tzu, *The Art of War*, we would find that the idea of treating workers with care and respect is not original.^{9,10} Management, to be effective, must also comprise leadership. Frederick Taylor, in his work *The Principles of Scientific Management*, brought important attention to the importance of managing the numbers but also took care to mention that the workers should earn a share of the prosperity resulting from improving the efficiency of their labors.¹¹ Henry Ford is remembered for his drive for the efficiencies of mass production and his groundbreaking \$5-a-day announcement in 1914 that aimed to enable his employees to acquire their own vehicles.¹² The worst and—unfortunately—most remembered aspects of using a moving production line and managing the numbers were first described graphically in 1906 by Upton Sinclair in his book *The Jungle*, about the meat-packing industry.¹³ Hounshell's work *From the American System to Mass Production 1800–1932* provides an excellent account of the development of these early manufacturing systems.¹⁴

The styles of management that developed fertilized the growth of the union movement and an inimical separation between workers and management. The unions did to some extent follow the pattern of the earlier guilds in providing qualification metrics and welfare for their members, but a principal role was as negotiators with management. A further unfortunate consequence was a proliferation of job descriptions that later inhibited cross-training, job sharing, and worker transfer. The leadership and management of any enterprise wishing to succeed must take note of the historical and linguistic baggage accompanying the words like management and workers and develop alternatives. Today, *associate* is a popular synonym for *employee* or *worker*.

In the second half of the last century a majority of the U.S. workforce enjoyed tremendous prosperity by comparison with workers in war-ravaged Europe and Asia. Nevertheless, there were strikes, hard negotiations, and, more latterly, waves of downsizings and reengineering causing lost jobs as foreign competitors grew more aggressive. However, the economy was generally robust, and some current opinions suggest that U.S. consumers were held to ransom

* Renowned for contributions to quality improvement worldwide and most especially in Japan, every text on quality offers ample descriptions of Deming's principles.

as both management and their workforce gained large pay and benefit packages. This was sustainable when the United States possessed a quasi-island economy, importing and exporting almost at will and with a positive balance of trade. As the economies, productivity, efficiency, and manufacturing prowess of competitor nations grew, conditions became arduous. Now major union tasks are to negotiate tiered pay scales, health care, pensions, and working or lay-off conditions. Since 1983 union membership has declined from 17.7 million, or 20.1% of a workforce of approximately 88 million, to 14.8 million, or 11.8% of a substantially larger 2011 workforce totaling 125.4 million.¹⁵

It is likely that union affiliations and power will continue to decrease. More workers are being empowered and given opportunities to become increasingly multiskilled. In fact, the workplace is forced to become much more collaborative, and “team” oriented. Additionally, the vision of lifelong employment—doing one task serving one enterprise—has faded as marketplace pressures together with technological change create a need for greater flexibility and faster responsiveness in the value chain from product/service concept out to customer satisfaction.

Traditionally, enterprises became accustomed to large hierarchical operations with relatively specialized division of labor and aggregation into functional groups for purposes of command, communication, control, and planning. These large and often “vertical” organizations took advantage of ideas of process simplification that were successful with lesser skilled labor. They enabled effective production and had few requirements for expanding the skill base of the employees. In unionized plants there was a profusion of job descriptions as well as levels and possibility of conflicts among workers with different crafts or unions. In a general sense, the skills became embedded in the tooling and in the fitters who set up the tools. This system was far from optimum, but based on theories of the time, skills available, social needs, and economics, it generated a reasonable level of prosperity. In a comparative sense, the long era of this style of mass production brought higher levels of wealth and prosperity to many more people and societies than any previous system.¹⁴

In the latter part of the twentieth century it became obvious that large hierarchical structures were a great hindrance to decision processes. There are many conflicts and appreciable difficulties in handling innovative ideas and change. Certain modifications were adapted from the military practice of creating special task forces, or teams with specific focused missions, operating outside the traditional reporting structures and management envelope. The “success” of task forces led to the adoption of many variations of matrix structures, disposing employees from different functional groupings into project- or program-focused teams. These matrix methods are contrasted with functional groupings in numerous treatises dealing with management.

A current example is the spectacular transformation at Ford from that of a confused chaotic episodic organization with many internal and warring fiefdoms described by author Bryce Hoffman in *New American Icon*. Hoffman relates new CEO Alan Mulally’s current efforts at Ford.¹⁶ Mulally has eliminated many reporting levels, introduced weekly and accurate status reporting sessions, and in flattening the structure has gained the confidence of a much reduced workforce, the board of directors, and the Ford family and their descendants, which is no mean feat. Admittedly, this is a work in progress; the marketplace, the tenuous global financial situations of 2013, and the ultracompetitive nature of the automotive industry will undoubtedly be factors influencing an enduring success. Nevertheless, the account of what Mulally achieved at Boeing and now initially at Ford describes some effective modern management principles.¹⁶

As mentioned above most large organizations are unavoidably dyslexic; they become bureaucratic and fossilized. Any organization eventually develops to preserve forms, stabilize activities, and provide secure protocols for our interpersonal behavior. Organizations of their nature inhibit change and restrict the development of ideas leading to continuous improvement. To be successful in the future, organizations must be structured with a recognition of the ineluctable life cycle of inception, growth, and maturation, with a, perhaps, evanescent stability

preceding the inevitable decline. A similar cycle is shared by every process, product, and individual associated with an enterprise, although with varying time constants. Organizations must be structured (and restructured) with a facility to accept and adapt to continuous and often unpredictable change.¹ Fresh paradigms must be evaluated and welcomed continually. There is need to create a pervasive awareness that stability is unwelcome.

In developing our ideal organization structure that is accepting of change and improvement, it must be recognized that the success of the earlier hierarchical pyramids was associated to a great extent with the collocation of individuals with similar affinities. Cross-disciplinary or matrixed cross-functional teams are a wonderful idea, but it is important to recognize that few individuals choose their career paths and disciplines by accident. These choices are related to their own social or psychological attributes. The most successful individuals, it can be assumed, are those who attain the closest match between their internal psyches and their professional activity. For example, there are appreciable differences in the communication and perceptual skills of many electrical and mechanical engineers. Such contrasts and potentials for conflict and team disruption become even greater as the needs of a team call for involvement from additional disciplines, such as accounting, economics, ergonomics, finance, industrial design, manufacturing engineering, marketing, materials management, safety, waste management, and the like. These interpersonal factors are exacerbated when different divisions of any large enterprise must collaborate or when international cultures are represented. All individuals have differing interpretations of the world as well as their own responsibilities to the enterprise and to the project at hand. The integration, management, and leadership of diverse multifunction teams require skills equal to those of the best counselors and therapists.¹⁷

4 ORGANIZATIONAL BEHAVIORS, CHANGE, AND SPORTS: FRUITLESS QUEST FOR STABILITY

It seems implicit in the human psyche that we assume tomorrow will be a close approximation of our “ordinary” yesterday. Both as individuals and as groups in organizations, we assume that “if only we can get over this workload hump, or this crisis, and past the next checkpoint and deadline, then we will enter a domain of calm and a plateau of stability.” In the main, our organization structures, measurements, and expectations are based on this idea that stability is an attainable and virtuous state. In the affairs of man this is patently untrue. At no time has history been free of change and of concerns for the unstable future. Explaining and forecasting this future occupy many economists. Kondratieff produced his ideas of waves following innovations or major changes in 1924. Joseph Schumpeter, expanding the initial idea that Werner Sombart (1913) derived from Marx’s *Das Kapital* (1863), further explained the idea that new methods or technologies resulted in the creative destruction of older systems.¹⁸

Notwithstanding these ideas about change, it is clear that from the earliest of times the human race has endeavored to organize itself to achieve surprise-free environments. We tend to gravitate to those groups that we know, where we will be safe, sheltered, understood, and free of surprises. In general, both individuals and organizations shun change. Enterprises create organizations to prosecute their objectives and to advance their interests. Every organization, if it embodies more than a few people, is compelled to develop bureaucratic structures to handle routine matters uniformly and expeditiously. Organizations of their nature strive to create surprise-free environments for their customers and employees. Thus, we see that people and the organizations in which they arrange themselves are highly change resistant.¹

Studies exist that demonstrate extraordinary productivity results when people are placed in self-managed teams with significant challenges in highly constrained environments. An idea and personnel are isolated and left alone and brilliance emerges, notwithstanding an awful environment and severe constraints. This has been called a mushroom effect because spores,

or ideas, are left in a dark corner on a pile of metaphorical horse manure and almost forgotten. There is substantial literature relating tales of bandit or pirate operations working against impossible deadlines with minimal resources, thereby becoming extraordinarily motivated and sometimes flouting the expectations of a mature parent organization. Stories of the success of small entrepreneurial endeavors abound, but there are many failures. Some of these projects are poorly structured but, nevertheless, succeeded as a result of the personalities of the leaders. Memorable examples have been excellently described by Kidder in *The Soul of a New Machine*, a book about the development of a new Data General computer model, and Guterl with his Apple Macintosh design case history.^{19,20} Subsequent technological transformations have been stimulated by variously charismatic leaders initially starting companies or projects with small footprints and few historical traditions as handicaps. These originators include, but are certainly not limited to, Google founders Sergey Brin and Larry Page, Bill Gates (Microsoft), Steve Jobs (reviving Apple), Mark Zuckerberg (Facebook), and Elon Musk (SpaceX, Tesla, and earlier PayPal), among others.

Many enterprises recognize that major improvements, such as accelerated new product development and introduction, require a different organization. They attempt to accomplish this by embedding specially assembled project groups within an existing but already archaic hierarchic framework. The transfer, or loan, of individuals with special skills into special quality circle task forces, early manufacturing involvement (EMI), or concurrent engineering teams is often an effective solution to overcome the dyslexic characteristics of a historic organization structure. However, it can be postulated that any success may be wholly due to the close attention that “special” projects receive from senior executives and is likely to be transient. It is difficult to evolve special teams into an ongoing search for continual improvement. It can be observed that these special high-profile teams lose their adrenalin fairly rapidly, and a string of me-too results follows. Ideally any major changes, new processes, or new product developments should be accompanied by a reconfigured organization. Special measures and personnel rotations are needed to ensure refreshment, revitalization, continual organizational evolution, and renewal.

When we compare practices in the arts and sports with those of industry, we can see many parallels. Clearly extraordinary performance can be generated by organizations that may be perceived as almost anarchist in character (cf. jazz groups). However, some form is detectable by the team members. Many leaders talk of teams and imply analogies with sports activities; others use the arts, and Drucker speaks of orchestral management.^{21,22} In many team sports the emphasis is often placed on moving a ball effectively. Aficionados of each different sport know exactly what is effective in their context. In most cases, the specialties of the players rest on either particular hitting skills or handling skills. In some cases, there are special positions on the field or pitch with a subsidiary requirement for either hitting or delivery. For the handlers, delivery becomes everything. They specialize; they practice; they examine every move in slow motion; they visit psychologists, chiropractors, and frequently specialist surgeons to improve and maintain their skills. They are rested, rotated, and measured with great refinement. Their rewards are public record, and they are accorded the esteem of their peers. Even with the star systems, most individuals and their management recognize the interdependencies of an effective team.

In team sports that do not involve a ball or puck, the measure of final excellence or speed may be easier, but integration of the individuals can be more difficult. Rowing, for example, requires great individual ability, but this is worthless in a four- or eight-person team unless the output of the whole team is synchronous. The bobsled event may look like the application of brute force with pure gravity, and the margins are remarkably tight. To the nonexpert, the contest results almost appear random. However, there is a regularity and consistency expressed in hundredths of seconds that demonstrates the excellence of the best teams. Measurements for attaining team excellence are demonstrably much more than just the assembly of the fastest pushers. The ability to think and act with one’s fellows and get onto the sled at the last possible

moment also plays a great part and cannot be measured by singular tests. However, the measure of integrated team performance is conclusive.

5 SYSTEM OF MEASUREMENT AND ORGANIZATION: STIMULATING CHANGE

Building on the sports analogy, an enterprise wishing to improve must consider itself as engaged in some cosmic league of global proportion. Although continuous improvement and high productivity are abstract concepts, they must be understood and defined in the context of the organization seeking to excel. There must be benchmarks; some “stake in the ground” must be established. A product cycle can be judged against historic comparisons or competitive benchmarks, and the time to initial generation of profits can be contrasted with earlier products. A higher productivity product cycle will reach the breakeven point faster and with less trauma within the organization. Institutional learning or human resource development should be an additional measure, as this has strong correlation with future prosperity.

Clearly, customers, shareholders, employees, and other stakeholders are continually measuring the attributes of the enterprise with which they are involved. The sum of these measures could be said to be the value placed on the enterprise by both the engaged communities and the stock market. This aggregate value is a composite measure of management competence, adherence to targets, efficiency of resource utilization, customer satisfaction, and product/process elegance. Elegance is a subjective measure that could be assessed from reviews of industry consultants, or experts. It may also be inferred from customer experiences, warranty claims, life-cycle costs, and level of engineering change orders, or equivalent measures in service industries. Since 1987 the extraordinarily successful implementation of the Malcolm Baldrige Awards demonstrates that it is possible and very worthwhile to make useful measurements of the many intangibles in business, health care, and educational environments.*

Such measures can readily be adapted for individuals and teams as well as organizations. Criteria for the Malcolm Baldrige awards are presented in Fig. 1.

Once there is a measure of the enterprise, it is relatively simple to decompose this and abstract a measure for every division, site, or department in the organization. This may well relate to long-term revenue projections, short-term profitability, or volumes, new-product introductions, market share, or global rankings; the organization measure adopted is a strategic issue for the enterprise. Any sports team or arts group possesses some intrinsic ability to judge its standing in whichever league it chooses to play. Ultimately, this becomes a numerical tabulation and is a measurement of organizational effectiveness in competing in the chosen market. The measurement intervals used must relate to the life cycle or time constants associated with the product cycles and the overall rate of change within the industry.

Further decomposition can be undertaken to evaluate each team and the individuals therein. Individuals making contributions to several teams will carry assigned proportions from every team evaluation. Individual evaluations (and rewards) should include recognition of all contributions to each team with which the individual was engaged. There should also be components acknowledging creativity, innovation, extraordinary contributions, an ability to integrate, and development of future potential. A valuable contribution to performance measurement can be gained by seeking reviews from the team colleagues, managers, and technical coordinators or leaders that work with the individual being assessed. There are a variety of ways to administer these 360° reviews and it is important that they are treated seriously and confidentially as a potential aid for improving performance. Each employee (or associate) may nominate colleagues for including in her/his survey with the concurrence of the primary supervisor. The review process must be based on data from several sources and should be dealt with one on

* The Baldrige Award scheme is administered by the U.S. Department of Commerce through the National Institute of Standards and Technology (NIST). Details are available at www.nist.gov/baldrige.

Malcolm Baldrige Award criteria for performance excellence.

1. Leadership—how senior executives guide the organization and how the organization addresses its responsibilities to the public and practices good citizenship.
2. Strategic planning—how the organization sets strategic directions and how it determines key action plans.
3. Customer and market focus—how the organization determines requirements and expectations of customers and markets; builds relationships with customers; and acquires, satisfies, and retains customers.
4. Measurement, analysis, and knowledge management—the management, effective use, analysis, and improvement of data and information to support key organization processes and the organization's performance management system.
5. Human resource focus—how the organization enables its workforce to develop its full potential and how the workforce is aligned with the organization's objectives.
6. Process management—aspects of how key production/delivery and support processes are designed, managed, and improved.
7. Business results—the organization's performance and improvement in its key business areas: customer satisfaction, financial and marketplace performance, human resources, supplier and partner performance, operational performance, and governance and social responsibility. The category also examines how the organization performs relative to competitors.

Figure 1 Malcolm Baldrige Award criteria for performance excellence.

one as a coaching session. There should be no surprises (or fear) because all contributors to a well-managed, continuously improving operation should have been encouraged to acquire superior levels of consciousness in their relationships with other team members and leadership. Measurement schemes must stimulate continuous lifelong learning and professional growth. After all, the human resources of any enterprise are avowedly the most potent and responsive resource available for enhancing quality, productivity, and continuous improvement.

In larger organizations during recent decades there has been sufficient turbulence, internal rearrangement, and reorganization, with reassignments to new programs such that hardly anyone had an opportunity to attain stability. Some of this churn was not productive for the enterprise overall, although there was appreciable, often involuntary, vitality added to the careers of affected personnel. Our new evaluation processes must recognize the life cycles of the organization, teams, and individuals. Change must be deliberate and planned. It should not necessarily be assumed that any individuals should stay with a project through the whole life cycle. There should be changes on some planned matrix, relating to the performance and developing (or declining) capabilities and interests of each employee, the needs of the project, and the requirements arising elsewhere within the organization. It is essential for the prosperity and success of the enterprise that any battles for resources, headcount, and budget allocation details between different departments, functions, and divisions are dealt with swiftly so that they do not impact morale and responsiveness. Musicians and athletes change teams or move on to different activities. Similar career styles in engineering should be anticipated, encouraged, and promoted by the measurement schemes adopted in all organizations that aim for continual improvement. There is need for circumspection when there are excellent contributions by departments, teams, and individuals to projects that fail or are canceled. Clearly, some rewards may be merited, but only if there was useful learning consistent with the longer term interests of the enterprise.

Organizational maturity implies a tendency toward a stability that can impede change and improvement. Therefore, it is essential to create measuring and management strategies that discourage the onset of maturity. There is a clear need for the stimulation and excitement occasioned by a degree of metastability. However, there is a contrasting need for security, stability, and confidence in the enterprise to enable creative individuals to interact in relatively nonthreatening environments. We are reminded of Deming's concern for the abolition of fear—this must be balanced by a strong touch of paranoia about competition, the onset of process or product obsolescence, changing technology, and other factors expressed so well by Grove.²³ There should be expanding horizons and opportunities for individuals within every section in the enterprise, accessible to all the employees. Total quality objectives, improvement, and high productivity can only be approached when all individuals gain in stature and opportunity as tasks are integrated or eliminated. In quasistable or service industries, there must be anticipation of new markets as resources are released by productivity improvements.

Organization structures and measurement intervals must relate directly to product/customer needs. Recognition of suitable organizational time constants is an essential concomitant to delivery of well-designed products into the marketplace, with a timely flow and continuous improvement. The management structure that is likely to evolve from the use of these types of measurement schemes will have some orchestral or sports characteristics. There will be teams, project leaders, specialists, conductors, coaches, and the inevitable front office. The relationships between different teams with alternate priorities may resemble that between chamber, woodwind, and string or jazz ensembles in our orchestra. The imposition of the rotation requirements, the time constants, will cause these almost cellular arrays to grow, modify, evolve, and shrink in organic fashion responding to the demands and pressures of an environment. The most responsive organization will accumulate skills and experience in the manner of some learning neural network, and an organization diagram may possess somewhat similar form.

6 COMPONENTS OF MANUFACTURING SYSTEM: SIMPLIFIED WAY OF LOOKING AT SYSTEM

The manufacturing system provides concept implementation from design through realization of a product and completion of the life cycle to satisfy the customer and society. The manufacturing system can be said to exist for generating wealth in a societal sense.² It is useful from a design, planning, and improvement viewpoint to break down the internal aspects of the manufacturing system by contemplating the interactions of six major components: materials, process, equipment, facilities, logistics, and people. These components and their integration form the system, and their organization is affected by factors external to the system.

The manufacturing system transforms materials into products and consumes material resources such as energy in doing this. There are also waste products and eventual recycling to consider. This component embraces all physical input to the system and resulting material outputs.

Materials are transformed by a process; this defines chemical, physical, mechanical, and thermal conditions and rates for transformations. If properly understood, the process component is amenable to application of computer technology for sensing, feedback, modeling, interpretation, and control.

The processes require equipment or tooling. The equipment must possess the capability for applying processes with appropriate precision on suitable volumes, or pieces, of material requiring transformation at the required rates. The equipment must be intrinsically safe, environmentally benign, and reliable. Today most equipment is electronically controlled, and there may be advantages gained by interfacing with other tools through a factory network to

facilitate communications. (Feedforward of process data can permit yield and quality enhancements in subsequent processes if they are designed to be adaptive.)

Process equipment requires an appropriate environment and services to maintain proper functionality; it may also be integrated with material handling systems and other pieces of equipment. There are special requirements for provision of utilities, contamination control, waste management, access for materials input, and output, which must be addressed under the category facilities.

These components are integrated and deployed by logistics. The logistics comprise product, process, and systems design data; forecasts; development schedules; materials management; accounting, business, financial, marketing, and distribution arrangements; maintenance; and service, including eventual recycling requirements. This component is information rich and of similar nature to process, only in a more macrosense. These are factors that are subject to change while designs are being carried out; they are also liable to suffer dramatic instabilities after the system is brought online. The logistics component comprises a most fruitful area for research and innovative strategies, which can be a significant commercial advantage over the systems of competitors. There are several notable enterprises, such as Amazon, Dell, Federal Express, Lands' End, and Walmart, whose core competencies are primarily logistical rather than focused on technological differentiation.²⁴ *Strategic Supply Management* by Trent deals comprehensively with logistics and management issues associated with supply chains.²⁵

The whole system requires operating agents or people. A system is dependent on people as employees, customers, stockholders, or owners; as suppliers or subcontractors; and as stakeholders residing in communities affected by the system. There are, again, many unpredictable factors involving all aspects of human behavior.

There are significant human resource, leadership, management, recognition, and reward issues internal to the system. These also become a reflection of the expectations of the external society that accommodates the system. All people variously seek stability with secure horizons and shelter from turbulent times; however, in the new industrial society there can be no stability. Stability means no growth—and eventual decline. There must be pervasive quest for continuous improvement with lifelong learning. Some social parity must be equally accessible to all who make contributions. These ideas raise questions with regard to equality of opportunities for contributing to increasingly technological endeavors. Drucker²⁶ postulated a population of knowledge workers in the 1994 Edwin L. Godkin lecture “Knowledge Work and Knowledge Society—The Social Transformations of this Century,” at Harvard. His model of the future is certainly credible, and it places heavy responsibility on educational systems to equip individuals for this future. Investment in human capital is an essential aspect of all future planning. All these matters come down to how whole societies are organized, how expectations are developed, and the development of concomitant reward structures. These factors have great impact on improvement efforts and productivity, and there are significant differences across different regions and cultures. The tasks of inspiring collaboration and continued workforce enthusiasm present greater problems than the tasks of acquiring and deploying available technologies.

Although the classification into six components aids the internal aspects of the design, many constraints to the processes and choices for the components and their integration derive from the relationship of the new system to its environment. These systems are not closed, and they are subject to perturbations that affect economies, social groups, nations, and continents.

7 IMPROVEMENT, PROBLEM SOLVING, AND SYSTEMS DESIGN: ALL-EMBRACING RECYCLING, REPEATING, SPIRALING CREATIVE PROCESS

When improving and reconfiguring the manufacturing system, it is advisable to have a final future vision in mind. The characteristics of a globally ideal future manufacturing system must

be founded on sound principles of thermodynamics and design. Entropy conservation and minimization of trauma must be the governing rules, both for systems design and for associated organizational and social structures.² Significant emphasis must be given to quality of working life and conservation of resources. For such systems to prevail and be successful, environmentally acceptable, and sustainable, there must be recognition of the global commons, as espoused by Hardin,²⁷ the Greenpeace organization, and green enthusiasts. The systems must aim to be environmentally benign while providing useful products that satisfy human needs and solve human problems, meanwhile affording employment with wealth generation for the host communities and all stakeholders. To meet the competition, the systems must be able to handle frequently changing customer needs. This calls for fast design cycles, minimum inventories, and short cycle times to afford maximum flexibility and responsiveness at least cost.

The improvement, problem solving, and design activity must recognize responsibility for the whole system whereby a design is to be realized; design is holistic and must be total. It is not reasonable to design or improve products, or processes, independently of the system for realization and eventual revenue generation. Equally so, the whole manufacturing system must be consciously integrated with the needs of the enterprise, customers, and host communities. It should be noted that few products are everlasting, and neither are the organizations that strive to produce them. Organizations and their structures must be designed so that they adapt with comparable life cycles to the products that they aim to generate.

Any improvement program must be regarded as a pervasive activity embracing such divisions of labor as research, development, process planning, manufacturing, assembly, packaging, distribution, and marketing and include an appreciation for integrating the activity with the whole environment in which it will be implemented. There must be a thorough consciousness of all the likely interactions, both internal and external to the enterprise. Because this can become such a vast activity, it becomes a problem how it may be best organized for outcomes with the least trauma (and delay). There are now systems and software available for life-cycle management (LCM); these require large-capacity servers that may be beyond the ambitions of smaller enterprises. However, smaller business operations can now “rent” access to powerful servers and appropriate software from major corporate subcontractors and appropriate software from major corporate vendors. “Cloud” computing is a growing and increasingly popular strategy whereby organizations can use the Web for fast access to their information technology needs with a wide variety of mobile systems at many different and frequently global sites.²⁸ Collaborations in the life-cycle field (LCM) between industry, universities, and research centers are being stimulated through governmentally funded centers.²⁹

Designing any improvement activity must involve planning, interpretation of needs, assessment and ordering of priorities, and a definition and selection from choices. There are measurements, and some degree of organization of resources is implicit. Design is an art of selecting and integrating resources using diverse tactics to address problems with consciously optimized degrees of success. It is important to gain a full appreciation of the problem; today this is emphasized as paying attention to the needs of the customer. It should be noted that future manufacturing systems will have many customers—and not just those purchasing the products that are generated. To some extent, all those involved with and affected by the systems should be regarded as customers who must be satisfied. There are both internal and external customers, the next worker down the line, or the assembly operations across the country or ocean and then the end-use purchaser. This is consistent with the most recent thrusts emphasizing quality. The measurements of success may be objective (such as revenue or profit, increased throughput, higher quality) or subjective (such as elegance). In general, history shows that elegant but simple and economical solutions will deliver satisfaction.

By considering customers and the problem definition or statement of requirements, possible measurement strategies can be derived. The idea of success affords the converse opportunity of failure and implies a gradation of performance levels. Problems can be defined (however metaphysical and obscure), and the level of success can be estimated. The measurements may

be totally subjective, absolutely commercial, or physical, like durability, size, weight, and so on. Nevertheless, once there is a specified problem and an indication of measurements for a successful solution, then problem-solving design activities can proceed. Brainstorming through this matrix will result in an improved problem definition and a superior measurement structure. Later, this measurement structure will support the evolution of organizational and administrative arrangements, resource allocation, scheduling, and so on. There are a wide range of quality tools that may be applied to aid the analysis. Most of these have greatest value for mediating discussions and interactions among the improvement team. For smaller scale matters Pareto plots and Ishikawa, or fishbone, diagrams may be sufficient.³⁰ At a more complex level quality function deployment (QFD) (otherwise known as house of quality) techniques can be valuable or Taguchi principles can be applied to analyze system noise.³⁰ In order to iron out problems a Six Sigma approach may be effective.³¹ Many of these tools are exemplified in the vaunted Toyota Production System.³² Smoothness and placid but rapid flow without churn are good indicators of design solutions that are likely to lead to success.

At this point, it is important to gain an overall appreciation for the whole environment in which the problem of improvement is being addressed. The environment can be considered as that which cannot be changed. It is there, it is unavoidable, and it must be recognized and dealt with while undertaking the design. Next, there must be an appraisal of the schedule and resources, both material and personnel. These are all primary regulators of the quality levels attainable for the results of the project and have a substantial impact on final costs and eventual consequences. Meticulous attention to early organizational details, responsibilities, and communications ensures cost-efficient decisions as a project accelerates and as the rates of investment and sensitivities to risks increase. These considerations are not necessarily prescriptive, serial, or sequential, and many may be revisited repeatedly as the project progresses. The closest analogy for these procedures is to the helical design of the chambered nautilus. There is spiraling recycling of problem-solving processes with continual accretion, growth, and accumulation of learning as the final solution is approached.

Design or problem solving is like a journey and, just as there are many adequate alternate routes to a destination, there is a possibility of many different improvements or implementation schemes of equivalent merit. If the measurement system does not give a clear answer, then some measurement at a deeper level should be developed. The measurements should be as unambiguous as possible or there is danger to the morale of the team. There must be single choices for serious focus and further development or the explosion of options becomes too large to handle. Something that can be of assistance here is a search for analogs, a comparison with other systems, benchmarking, and analysis of the competition. Aspects that may be intangible from a viewpoint of strict functionality can have valuable impact in terms of brand, or corporate identification, ease of customer association, and so on. Such factors are all part of the team responsibility, and ultimately they can be measured, although subjectively. The process must proceed simultaneously (concurrently) with the development of any necessary changes in manufacturing system infrastructure. Additionally, designers should contemplate a risk analysis with best- and worst-case scenarios to cover either phenomenal success at start-up or utter failure. It is also wise to assess market volatilities and dependencies on unforeseen influences and competitive responses. These can range, for example, from drastic economic shifts due to oil embargos and energy crises, including carbon taxes through environmental regulation changes eliminating materials and processes, which could damage the ozone layer or be found to be toxic.

8 WORKFORCE CONSIDERATIONS: SOCIAL ENGINEERING, THE DIFFICULT PART

A key requirement of a system that desires continuous improvement is that the workforce is empowered and capable—that is, encouraged by continuous learning and with adaptability and enthusiasm for change. Here the ideas of Deming and other quality experts

are indispensable.^{9,30*} The ideas of total quality management (TQM), quality circles, and self-directed teams are valuable tools for operational improvement. In the case of Six Sigma implementation, it is a requirement that the procedures are learned and introduced by senior management. Each management level is required to be fully engaged and to collaborate in the training of the workforce. Several levels of accomplishment and attainment are recognized by colored belts (a judo analogy).³¹

Empowerment has many faces and can be very threatening to established bureaucracies. Teams must be empowered if they are to operate effectively, and they must be accorded authority to match the responsibilities of the problem(s) they have selected or been assigned. It is therefore mandatory to provide appropriate education and training opportunities for the whole workforce to ensure development of capabilities matching these responsibilities. Thoroughgoing empowerment should flatten organizations, eliminating many levels of management. Faster decision making and on-the-spot improvements enhance enthusiasm and participation of the workforce. Success engenders success. Increasing efficiencies are accompanied by quality improvements, reduced costs, improved throughput, and higher output. There are thereby opportunities for reducing manpower while maintaining steady production levels, and alternatively pricing structures can be changed aiming to deliver greater volumes and increase market share. Ideally, surplus manpower could be diverted to the development and introduction of new products. More customarily we hear phrases like downsizing or right sizing and reengineering. The improvement ideas just given are as attractive as they are simple, but thorough implementation is a severe test for the leadership and management of any organization that genuinely wants to promote change and improvement. To maintain these ideas requires the dedication and concentration usually reserved for sports teams and their coaches. In sports there is continual measurement, evaluation, rotation, and renewals. This will undoubtedly apply in successful industries in the future, and societies may well be compelled to adapt to continually changing and impermanent employment prospects.

There are two main motivations for workforce reduction, the initial one being cost reduction and the other being productivity improvements. Less labor reduces total costs, but recent analysis shows that direct labor costs rarely exceed 10% of the cost of manufacturing. Thus, any reduction in labor costs has only fractional impact; this measure similarly implies that relocation to countries with cheaper labor may be a false economy. The other, more persuasive reasons for involving fewer people all relate to productivity: quality, throughput, cycle times, flexibility, and responsiveness. The only work that should be done is that which adds value for the customer or ensures learning and future process improvement. When value engineering or reengineering is applied to customary organizations, it is found that there are many people keeping each other busy performing obsolescent tasks. Although technology is occasionally misapplied or is overly sophisticated, when used intelligently it enables workforce reduction and productivity improvements that afford cost–performance benefits. There are also opportunities to improve the quality of working life as a corollary.

As the workforce is reduced, the remaining employees become capable of working better, faster, and more effectively as tasks are redefined. The rewards and recognition for tasks well done are easier to determine. The elimination of layers of management, supervisors, and other titles boosts morale and improves communications, team spirit, and the rate and quality of results. However, depending on the levels of threat that created the need for change, it may take some time and care to raise morale. Simultaneously, individual responsibility levels and stress may increase as the nature of work and contribution in the workplace change totally. Workers deserve a level of security, support, and confidence if they are to take risks in making decisions for themselves. Further education and training become continuing life-long requirements. Often, the entire involvement with work may be forced to become a stressful integral

* See footnote on Baldrige Award scheme in Section 5.

part of the life of each individual. As in preindustrial times, the needs of society for production become inseparable from the life in the community, although machines and the intensity of international competition dictate what may be a less optional and crueler pace.

It is easiest to introduce new systems of these types as response to severe external threats or in times of great and imposed changes. There is instability, fear, and paranoia and, hence, a great appetite for new solutions. However, this still requires leadership, courage, and determination. Additionally, threats of bankruptcy or similar trauma, for example, may make conservation of resources a more urgent priority. Thus, it may be more vital to deal with the threat immediately to gain breathing space for later improvements and reorganizations. Gerstner gives an excellent account of his early days at IBM when securing some stability was more important than spending time developing a vision.³³ Nevertheless, his confident actions suggest that he did have his own nascent if selfish vision! An entirely new organization and approach with a clean slate stand the best chance for survival and prosperity. Due to prevailing conditions, Gerstner was compelled to transform the IBM organization and restore profitability somewhat stealthily.

Several methods have been described for developing team spirit and breaking down disciplinary and professional barriers that inhibit effective integration.³⁴ The General Motors Saturn Project broke new ground in organization structure and development of employee commitment.³⁵ The project started with a small, carefully selected cadre of employees, future associates, representing all levels of workers. This team was given responsibility of working with the architects and factory designers to resolve issues of equipment and facilities layout, break area, and restroom locations. They were given coaching sessions to introduce architectural concepts and notions of scale and space, safety and OSHA requirements, etc., so everyone was on the same page. The initial team was then charged with interviewing and recommending hiring of their future colleagues. When Saturn vehicles were first introduced, they were a well-received and successful team-based product. There were many new and different organizational features that carried through from design right onto the floor of the showroom. Customer satisfaction was high, for a time sales were limited by factory capacity problems, and expansion was debated. Subsequently, and most unfortunately, this grand idealistic experiment was found to be unaffordable, and Saturn was eventually folded back into a struggling GM organization. Notwithstanding significant cutbacks, reorganizations, factory closings, and workforce reductions, GM was compelled to seek bankruptcy protection in 2009 and a much debated government “bail-out.”

9 ENVIRONMENTAL CONSCIOUSNESS: MANUFACTURING EMBEDDED IN SOCIETY

It is clear that manufacturing has changed profoundly in the last few decades. Formerly manufacturing more closely approached the model developed from the blacksmith pounding hot metals on an anvil. Ultimately, these and the accompanying assembly processes were automated and became the mass production manufacturing lines of the last century. Now it is reasonable to regard manufacturing more holistically. The success of the enterprise requires that management satisfy the demands of the marketplace and meet the competition. These objectives no longer mean simply the generation of revenues, profits, minimized costs, and a stream of new products, with ever better and more comprehensive services to satisfy the customers (and stockholders). They also require sustainability with respect to the environment. Both require respect for the long-term needs of the global community. A manufacturing system can be defined as a means of transforming resources, including materials and energy, into products and/or services to satisfy customer needs and concurrently generate wealth for society without trauma and waste.²

9.1 Sustainability

The definition of sustain is to keep in existence—maintain.³⁶ Today, increasing attention is being paid to the whole concept of sustainability. In earlier times there was little thought given to the ideas of life-cycle engineering. Products were designed for an estimated, or forecast, life span, and there was no consideration of subsequent retirement, reclamation, recycling, or eventual disposal. The idea that product design must cater for the whole life cycle of any product from concept to end of life and safe disposal or reconfiguration for further use is relatively new. This idea also encompasses the safe disposal and/or reuse of process residues and elimination or significant reduction of emissions into the atmosphere or into the environment, ground, and water. In a manufacturing systems context, sustainability has the implication that any system should be designed and implemented according to certain principles with no negative effects, present or future.² Additionally, the products should be as benign as possible in use, and they should be readily capable of reuse, renewal, reclamation, recycling, or safe conversion and disposal. If these ideas had even been contemplated a couple of centuries, or even longer, ago, the Industrial Revolution would have been stillborn. The countries and economies that we now know as the developed world would still be reliant upon agriculture and crafts. The movement from agrarian and peasant-based economies to mercantile capitalism and globalization would have been slowed appreciably.

Aspects of design for sustainability are now receiving wide attention. “Cradle to grave” phraseology is accompanying some discussions on new products and systems. Here it is appropriate to encapsulate the four key strategies defined by Hawken, Lovins, and others.^{37,38}

Effective Use of Resources

Hawken and Lovins espouse amplification of efforts to use all resources more effectively.^{37,38} The major and high-priority aims are for radical resource productivity, reduced rates of depletion, lower pollution, and creation of jobs. The metrics implied are not only the immediate bottom line but also the generation of wealth for all stakeholders and the entire host community without traumas (present or future). This is consistent with precepts discussed in publications by this author.

Biomimicry

A second strategy is that of biomimicry. Here the aim is to copy natural systems of biological design and reuse (phytoremediation). In the forest, trees grow, mature, weaken, fall down, and rot, and the decomposition gives rise to many reproducing and multiplying organisms. Ultimately, more trees sprout up in a repeating and enduring sustainable cycle. It is worthy of recall that the first notions of sustainability were associated with crop rotation in medieval agricultural practice. Wilson, in *The Future of Life*, commends a case study in Guatemala where a small local population was enabled to live sustainably by marketing natural products from a stable rainforest, as opposed to selling the timber and creating farms.³⁹ Ben & Jerry’s ice cream company originated with these kinds of principles. Its product remains moderately successful among a growing group of other environmentally conscious offerings. Its practices follow the concept of the Commons espoused by Hardin in 1986.²⁷ This is in marked contrast to U.S. agribusiness practices in the fast-food supply chain for beef, chickens, hogs, and potatoes described by Schlosser.⁴⁰ The idea of encouraging small, flexible, individually oriented and entrepreneurial operations is intrinsically attractive. For people with the right skills it is possible to launch and then expand uniquely innovative firms using the Web. Nanotechnologies adopt a somewhat parallel approach in that atoms or molecules are experimentally configured into the form desired. Recently announced processes involve the precise patterning of an appropriate surface and atoms, or molecules, are then caused to deposit or grow preferentially in the presence of a catalyst to produce specifically oriented arrays with extraordinary properties. These

extremely small-scale processes are proposed for limited numbers of new and future products. However, although these processes grow the product in a quasi-organic sense, at this point they do not appear to be truly sustainable and there may be toxicity exposures. Meanwhile, on the human level, organizational analogies based on studying the behavior of ants and bees are being discussed.

Extending Manufacturers' Responsibility

The third strategy reminds us that we exist in a service and flow economy. This revalues relationships between commercial operations and their customers. The manufacturer's or product originator's task is not completed when the customer walks out of the store, showroom, or sales office; there are responsibilities beyond the instant exchange of cash. This aligns with the idea of extended products, which implies taking a lifetime product support perspective and includes all services to support the product in addition to the manufacturing of the product itself, and then retirement of the product. This may require that intelligence be embedded in the product. Accompanying this is the idea of providing an on-going relationship that gives customers solutions, experiences, and delight. There is greater emphasis on agility and responsiveness rather than on the product itself.^{41,42} Tangible and intangible products and services are included. It is a novel challenge to develop support structures for these extended products.

Conservation and Restoration

The fourth principle emphasizes the importance of conservation of everything and restoration of the original status wherever possible.

9.2 Principles for Environmentally Conscious Design

There are several corollaries for systems design that expand on these strategies, not necessarily in rank order. There is some inevitable overlap and redundancy:

1. There are no negative effects on the environment.
2. Conservation, elimination of waste or scrap, capability for reuse or remanufacture—product can readily be made equivalent to new (ETN).
3. Products or services are benign in use.
4. System should accommodate global concerns for energy, fuel, natural resources, and degradation—preserves balance.
5. The future system life cycle is considered (cradle to grave).
6. Quality of life of users (a “necessity factor”) is considered.
7. The development and future horizons of the workforce are important.
8. Ergonomics and health concerns are factors.
9. The effect of the system on the host communities must be determined.
10. Local, regional, national, and global ecological, economic, and financial matters are considered. The design embraces global warming, pollution, entrained diseases, and other factors that may affect the planet.
11. System should be based on smooth flows—evolutionary, organic, and nondisruptive.
12. The design must respect people as individuals—customers, employees, stakeholders.
13. There must be only positive long-term impacts.
14. The system overall must improve the common weal.

10 IMPLEMENTATION: CONSIDERATIONS AND EXAMPLES FOR COMPANIES OF ALL SIZES

Probably at no previous moment in world history have commerce and industry been so complex, extensive, and globally interrelated. In the twenty-first century, almost all of our commercial activities have global connotations and are increasingly digitally based. The growth of information technology (IT) has accelerated a shift from the simple exchange of cash for products or services to entire manufacturing or commercial systems that market solutions enabling customer success, provide experiences, and build continuing relationships. Process execution is critical; the process is continual and not intermittent, as it was in the past with discrete transactions. The provider entity must learn metaphorically to walk continually “in the moccasins” of the customer. Customer delight is the major metric for attention, but at the same time the needs of the community must be recognized. Strategies that emphasize customer success are imperative; these may often dictate collaboration with former and even present competitors.⁴³ In the twenty-first century successful enterprises must recognize the power of their customers (and the market) and must endeavor to establish long-term relationships based on providing solutions, service, and a successful ongoing experience that replaces the earlier tradition of product delivery and cash transfer.³⁵

10.1 Vertical Integration

Digital business (or e-business³³), together with “cloud computing” and social networks, affords many new ways of operation.²⁸ The factor of speed to customer delight challenges industry strategists to incorporate many new approaches and activities.

The most successful enterprises have reorganized and refocused their activities. Vertical integration with control of all the materials, processes, and operations that go into producing a product is no longer preferred. This was the model adopted long ago by Henry Ford at his Rouge plant.¹⁴ It is now important to focus on core competencies, activities that depend on special skills, or advantages that the enterprise possesses or has developed. These are the processes that represent the intellectual property or crown jewels of the organization and generally provide the best opportunities for revenue maximization. Subsidiary operations can be delegated to specialist subcontractors. (Referring back to our athletic models, it can be observed that all-rounders do not often win Olympic medals.) These tiers of subcontractors are organized into supply chains and are managed electronically. Excellence in communications, delivery, reliability, and quality are all key requirements to support just-in-time (JIT) manufacturing systems.

There are risks and some disadvantages with the dependencies occasioned by lengthy supply chains. Larger enterprises frequently insist that their suppliers build factories and/or warehouses nearby final assembly plants. Some enterprises have specialized contractors designing, developing, and manufacturing substantial proportions of major components. Boeing is a key exponent of this technique. Its website lists worldwide contributors to the 787 Dreamliner project (<http://www.boeing.com/commercial/787family/>). There are multiple reasons for delegating production activities; expertise and focus are primary. Also there is cost; large organizations may carry very high overheads due to research and development, infrastructure, and legacy costs. When these are written off against the manufacture of noncore components, parts, or sub-assemblies, then the raw cost of these items becomes unsupportable. For example, in the early 1980s when IBM introduced its personal computer, the internal pricing of its world-class hard drives, memory, and microprocessor chips would have made realistic market pricing impractical. IBM relied on vendors for all these components, the plastic housing, and peripherals.

10.2 Real-World Examples

Dell affords one of the best models for reliance on supply chains. Its core competency is giving customers exactly what they wish for just as quickly as Dell can have it assembled and

Table 1 Toyota Precepts

-
1. Fostering an atmosphere of continuous improvement and learning
 2. Satisfying customers (and eliminating waste at the same time)
 3. Getting quality right the first time
 4. Grooming leaders from within rather than recruiting them from the outside
 5. Teaching all employees to become problem solvers
 6. Growing together with suppliers and partners for mutual benefit
-

Source: From Ref. 32.

shipped. It uses an array of suppliers to fulfill carefully figured inventory needs for just a few days production. Computers and peripherals are custom-built through final assembly in direct response to orders. Some bare chassis may receive standard boards and components, and special models may be finished for promotional uses. The history of its initial direct-to-customer build-to-order PC process is described in a book by Michael Dell.²⁴ As a result of other perturbations and strong competition in this business, Dell, once a leader, is now third in global market share, with 12.1%, by comparison with HP and Lenovo, with 17.3 and 13.0%, respectively, for 2011. Dell reported revenues of \$62.07 billion in February of 2012.

There are three Golden Rules: (1) disdain inventory, (2) always listen to the customer, and (3) never sell indirect. Dell also segments so that each business unit stays “small and focused on the needs of specific sets of customers.”²⁴ Business information and plans are shared with all employees. Dell is virtually integrated through the purchasing arms of its large account customers. It continually strives to step back to examine the whole context and environment and avoid “breathing your own exhaust.”²⁴

One can hardly discuss manufacturing system improvement without mentioning Toyota (see Table 1). The Toyota Production System (TPS) has been discussed and written about at length, and many Japanese phrases have passed into the manufacturing lexicon. Kaizen is a watchword, and the principles of lean manufacturing are covered comprehensively in other texts. Expressing it simply, lean manufacturing involves only doing things that add value for the eventual customer, eliminating waste in all its forms, and not building inventory between operations or at the end of the manufacturing line. Theoretically, a request by a customer must pull products out of final assembly or processing, and once that product is on the way, then the upstream processes must be instantly ready to fill the void (JIT). In the idealized case, every internal or external unit in the supply chain must respond identically. The flow may be controlled and maintained by exchange of signals, or kanbans, up and down the delivery chain.³²

When we think of manufacturing systems, we tend to think of Boeing, Ford, General Motors, IBM, Johnson & Johnson, Pfizer, and similar giant enterprises with multiple and globally distributed sites. Their actions and strategies are often drivers of the economy, but they would be powerless without the contributions, endeavors, and support of the smaller units tiered in the supply chain. In fact, a very large proportion of all manufacturing is undertaken by much smaller firms. Analysis of data from the Bureau of Labor Statistics for 2008 indicates there are almost 6 million companies with less than 500 employees comprising some 49% of the whole workforce of approximately 121 million.* When it comes to the large companies, there are just short of 1000 firms (981) in the United States with more than 10,000 employees that provide just over 33 million jobs for 27% of the workforce.

More than 5 million firms have fewer than 20 employees accounting for just over 21 million employees (18%).

* Multiple U.S. government sites offer many statistics, including www.bls.gov. The data shown here are from www.census.gov/econ/smallbus.html.

The importance and significance of the contributions of these many smaller companies may often be neglected or at best underestimated in learned treatises. There may be implicit assumptions that the latest leading edge software, systems, and associated technologies are available, affordable, and accessible. Smaller firms suffer difficulties in this connection and may be forced into bidding wars with OEMs that further constrict their resources and development budgets (if they exist at all!). On the other hand, they have the potential advantages of less bureaucracy, faster decision making, and possibility of greater flexibility with responsiveness.

The case of Liberty Brass in New York City is a model of the advantages that accrue to a smaller nimble company with courage and imagination.⁴⁴ This is a 40-employee machine shop with adroit equipment choices and reduced set-up times to turn around smaller and special orders responsively. It improved planning and estimating using the latest software so that quotes and eventual deliveries were accelerated and outsourced suitable jobs to specialists. Essentially, it focused on core competencies and developing relationships for speedy customer service. Its prices are higher than Chinese suppliers, but its quality, reliability, responsiveness, and lower shipping costs gain Liberty Brass repeat orders.

In a similar and more recent case, Watermark Designs, a Brooklyn (New York) operation with 45 employees, were able to invest in a 3D rapid prototyping unit and are thriving delivering high-quality custom-designed plumbing fixtures for luxury condominiums and hotels in Shanghai, Macau, and Hong Kong.⁴⁵ Export revenues from companies in the New York metropolitan area totaled \$105 billion in 2011.⁴⁵

10.3 Education Programs

Improvement systems cannot be implemented without education and training. Any program must have wholehearted commitment and demonstrated support by senior management (as required by Six Sigma). There must be not only commitment but also time spent, either in welcoming employee students to the first sessions and introducing the objectives or as a visiting speaker for an executive overview, luncheon, or coffee break. The program must be accessible to all qualified or selected employees without undue conflict with regular responsibilities. There should not be any work penalty associated with attendance. All participation for work-related technical vitality and improvement programs should be on the clock and during working hours unless there are special exceptions. When Harley-Davidson wished to transform its operations, its analysis showed that it needed to hire a few additional people so that everyone could be scheduled for up to one hour per week of education or training time.

The long-range objectives to be accomplished must be clearly enunciated.

Preferably, common administrative phrases and buzz words should be avoided in all announcements and program descriptions. Words worth avoiding include business conduct, continuous improvement, diversity, ethics, lean, quality or quality circles, safety, sexual harassment, and teamwork—these are important topics that should be embedded into programs with newer, different, and more invigorating or imaginative titles.

Ideally, educational activities should be accessible to the whole workforce, including contract workers; everyone in the workplace should be on the same playing field. Special programs are appropriate to certain workers, levels, and technical areas. There should be some base program for everyone on a rotating basis, with time included in regular work assignments (maybe up to 40 min a week or every couple of weeks) for training. Consistency and regularity are important; employee stimulation must endure after the first banners in the cafeteria fade. Training should cover vital topics such as quality tools, assessments and appraisals, career planning, latest trends in teamwork, personal development, and investments and retirement. State of business sessions with local executives or politicians are a useful change of pace. Special cross-training needs can be accommodated using in-plant experts. Mixing individuals from

Table 2 Essential Ingredients for Improvement

-
1. Vision: aspirations and dreams
 2. Mission, goals, and objectives (metrics)
 3. Personnel and their motivation (empowerment and teams)
 4. Strategies and tactics (core competencies, cooperation, and partnerships)
 5. Customer focus (experiences, relationships, and solutions)
 6. Communicate and listen to all stakeholders
 7. Priorities and resources: consistency and persistence
-

different departments and functions in every group and having employees introduce their new-found friends encourage interactions. The program must be perceived as worthwhile and not a timesink. Sessions should be short, frequent, and regular, with occasional promotional items, badges, ribbons, or certificates. Above all, they must not be boring. The occasional use of an outside facilitator and DVDs or tapes by speakers such as John Cleese or Tom Peters can be very worthwhile. There are also a wide variety of materials available on the Web through TED (Technology, Entertainment, Design, www.ted.com) and YouTube (www.youtube.com) that are of very high quality. Make sure there is lots of action, vitality, and stimulation with immediate value, discussion time, coffee breaks, graduation breakfasts, and similar events. Finally, measure everything to support continuous improvement of the program. Useful books for handouts are available in bulk from suppliers such as www.goalqpc.com, a good source for memory joggers and Six Sigma materials, and Price Pritchett at <http://topics.practical.org/browse/PricePritchett>.

10.4 Measuring Results

Successful implementation of improvements can be verified with traditional production metrics. The most convincing metrics are with regard to costs. Improvements and quality are not exactly free. Investment of resources and time is required to accompany, continue, and maintain appropriate schemes. Customary accounting methods can needlessly introduce deceptive chimera into cost analysis. For example, a small, extraordinarily responsive company making custom products to order is likely to have low inventories, a small amount of work in progress, and a limited backlog. Notwithstanding excellent revenues and decent profits, if such a company wishes to expand and seeks additional capital, it possesses few assets and apparently limited prospects. Any banker would be skeptical of approving loans for equipment that would facilitate growth, increased market share, or entry into a related market. Inventory and materials or partially finished goods in the pipeline may be counted traditionally as assets, whereas in the prevailing dynamic conditions they are a practical liability that must be sold off before newer, more current products can be introduced. In fact, to the loan officer a poorly managed operation loaded with inventory and work in progress (WIP) could be regarded more favorably than a bare-bones, fast-turnaround, highly responsive operation. As a result, when measuring the fiscal value of improvement schemes, it is vitally important to secure the costs and savings numbers that truly reflect the performance of the new or improved activities. This technique is known as activity-based costing. Once adopted, it leads directly to activity-based management (ABC/ABM).⁴⁶ Table 2 shows a list including some of the more intangible factors that should also be considered among the possible assets of an organization.

Costs are very sensitive to the methods used for sharing the overhead burden. The ABC analysis looks at the costs that are directly attributable to the component, part, or assembly being

manufactured—in other words, the direct cost of production. Analyses using this approach can reveal surprises when comparing manufacturing costs for specials against high-volume regular offerings. ABC analysis makes it possible to decide which of several ranges of products show the best cost-to-revenue ratios. Traditional accounting methodology does not do this effectively. Explorations of this type are essential when comparing the costs of a part made at the home location against the final delivered cost of the same part shipped from abroad. Companies must factor in costs of goods in the lengthy pipeline and possible risk exposures when deciding whether to make or buy.

11 A LOOK TO THE FUTURE

The ideal manufacturing system of the future, no matter how large, must be flexible and responsive and simulate the performance of a small operation. It must be focused on the best and most effective methods of bringing its core competencies to the service of customers and on maintaining relationships with these customers. Excellent relationships, collaborations, and partnerships must be established with subcontractors, suppliers, and associated groups for mutual benefit. Teams will comprise individuals within an enterprise and also employees of partner enterprises. Virtual and global relationships and dependencies will become customary, and even essential, as a response to unpredictable and urgent challenges and opportunities. The employees of manufacturing organizations that wish to survive and prosper in the future must be well-trained, knowledgeable, and empowered team players with excellent communication skills and empathy for different cultures.¹⁷ Management and continuous improvement of these operations call for vision, enthusiasm, and excellent leadership skills. Table 2 highlights important ingredients.

There are significant dichotomies as all these concepts are developed and integrated. There will always be economies of scale, but penalties of increased impedance, noise, and reduced quality will always accompany bureaucracies and size. By increasing degrees of order, it becomes easier to perform standard operations, but by employing total centralization, degrees of freedom dwindle, there is less variety, innovation is suppressed, boredom grows, and systems degrade.¹

This paradox applies universally and requires that leadership of large organizations acquires great skills of diplomacy and humility; very difficult and balanced judgments must be sold to diverse constituencies. Modern management requires walking a tightrope and balancing the priorities for survival and prosperity of the enterprise against the desires of the stakeholders and host communities. Some short-range suffering may be inevitable, as skills, abilities, needs, economics, and manufacturing strategies are rebalanced in response to global challenges. As Professor Joseph Stiglitz observed in relation to globalization trends and lean implementations, “Short term losers are concentrated, often very noisy, and their pain can be considerable ... The average worker in rich countries may actually be getting worse off, and there are probably more losers than winners There is need for a strong social safety net ... with more progressive taxation.”⁴⁷ Ideally, local rearrangements can be deployed in response to the most punishing inequities. More recently Stiglitz has authored a book discussing the growing gap between corporate leaders and all those in charge and the so-called middle class.⁴⁸

Re-education programs, community improvement projects, and expanded leisure opportunities, with adequate funding, are possible alternatives to early retirement or unemployment. It is important to remember that all resources, whether human or material, are finite and should not be wasted. All opportunities should be grasped effectively, and new means of securing revenues may require some creativity.

It is a dream that the manufacturing systems of the future will be not unlike a drive-up fast-food operation or the food court of a major shopping mall. There will be arrays of small factories that each can accommodate multiple needs but represent specialized sets of skills

(or core competencies) materials, processes, assembly and fabrication techniques, and finished items. There will also be series of agents, product integrators, advisors and consultants, or sales counselors, their function being to advise and act for customers as an interface with the various manufacturing facilities required to procure the desired custom product. Possibly all these functions could be handled through networks or catalogs in some virtual manner, as often already applies today. Customers will design, construct, buy, and equip their own houses, kitchens, computer systems, meals in restaurants, and other goods. Behind the retail suppliers are arrays of agents, distributors, wholesalers, and ranges and sizes of subcontractors and manufacturers. The future will be some spectrum of all these possibilities, with an extraordinarily wide range of time constants affecting the acceptance of new technologies across industries, regions, countries, and continents. Even in the presence of new technologies, there is little doubt that examples of past practice will continue to add value effectively in certain applications. There will always be the need for highly skilled craft workers to develop, fit, and maintain new tooling and, although overall the number of old-style tool-and-die makers has diminished, there will be niche areas for specialists in activities regarded as obsolescent.

It will always be prudent and economically reasonable to assemble many products in proximity to the final customers. This shortens delivery, reduces inventories, increases supplier responsiveness, and accelerates the concept-to-cash metric. Large enterprises are now reconfiguring their factories and replacing dedicated “hard” tooling with flexible tooling capable of fabricating a wider range of products. These factories are also being located proximate to centers of population nationally and worldwide. Where the large enterprises go, their suppliers follow. The former megafactories with several thousands of employees have been displaced by smaller distributed units that are simpler to manage and control. The existing paradigm of employing fewer people working “flexibly” for almost uncountable hours should be questioned. From a human resources aspect it is not sustainable or even socially efficient to overwork an employed few and not incorporate the un- or partially employed among the contributors to the prosperity of the whole community. Ideally, accumulating and increasing education, leisure, and recreation opportunities with reduction in hours worked by individuals would make for a more equitable society. More people working for fewer hours would expand markets for leisure and recreation opportunities (meanwhile reducing unemployment). In 1930 economist John Maynard Keynes suggested that within a century a 15-hour work week would be sufficient to satisfy personal, business, and industrial needs. In the same year the Kellogg Company introduced a 30-hour work week that endured from 1930 until the 1980s.⁴⁹ There are some companies today with practices along these lines; flex-time, working at home, having time allocated for independent activity (but with possibly company-related results); this encourages work to become somewhat like a hobby with compensation and benefits. It takes unusual financial and intellectual property strengths and compatible leadership/workforce relationships to engineer conditions of this nature.

Triumphing over global competition may result in a Pyrrhic victory and one cannot but envy the kingdom of Bhutan that esteems gross national happiness (GNH) as a more culturally acceptable index than the gross domestic product (GDP). In our postulated highly competitive society some individuals work probably harder and with more stress than they prefer and many are unemployed (or underemployed with occasional minimum wage tasks) with little opportunity for gaining creative satisfactions and no time or energy to do so. Meanwhile the salary and wealth differentials between the corner office holders and the average worker continue to expand. In order to implement continuous improvement regimes successfully a whole company and ultimately the whole of society need to become much more cohesive and similar to the most exceptional idealized top performing athletic or sports team. Peter Diamandis presents a view of possible futures that could embody some of these notions.⁵⁰

These trends must be accompanied by an increased focus on the continuing education and technical vitality of the workforce, but we must anticipate the future eagerly. Greater numbers

of distributed factories serving customers will entrain more jobs, and there will also be service, maintenance, education, and health care needs to be satisfied, thus generating prosperity and consequently more customers. As the population ages, provided they are allowed to continue earning, and a community remains prosperous, there will be an increase in discretionary expenditures for recreation and travel and again more jobs. There is need for global acceptance of the cycle that jobs engender customers and that prosperous customers fund more jobs, and then those jobs empower more potential customers. As customers of these systems ourselves, we must consider whether these notions are satisfying or efficient. Such questions can only be resolved by the system of measurements that must be implicitly agreed upon as we almost accidentally embark on improvements socially and professionally. The important thing is to start the journey knowingly. We may enjoy and learn from our nostalgia for the past, but we must anticipate the future eagerly.

Author Note

This chapter owes much to substantially amended, modified, and updated rewritten materials extracted from prior works by the author. The cooperation of the publishers of these items is gratefully acknowledged.

Author's Works

- “Design: Organization and Measurement,” in *Handbook of Design Management*, Basel Blackwell, Oxford, 1990, pp. 155–166.
- “Management Structures for Realization of High Productivity,” in *Proceedings of the International Conference on Productivity and Quality Research (ICPQR-93)*, Vol. 1, Industrial Engineering and Management Press (IIE), Norcross, GA, 1993, pp. 237–246.
- “Globally Ideal Manufacturing Systems: Characteristics & Requirements,” in *Flexible Automation and Integrated Manufacturing (FAIM'93) Conference Proceedings, Part II, Computer Integrated Manufacturing Systems, MIA: CIM Strategy*, CRC Press, Boca Raton, FL, 1993. pp. 67–77,
- “An Integrated Design Strategy for Future Manufacturing Systems,” *J. Manufacturing Syst.*, **15**(1), 52–61, January 1996.
- “Management Strategies for Sustainable Manufacturing,” in *Proceedings of the Seventh International Pacific Conference on Manufacturing*, Vol. 1, Bangkok, Thailand, November 27–29, 200, pp. 13–22.
- “The Evolving Production Enterprise,” in *Proceedings of the 18th International Conference on Production Research*, Salerno, Italy, International Foundation for Production Research (CD), 2005.
- “Manufacturing: The Future,” paper presented at 16th National Manufacturing Week, Chicago, Session 2D21, Reed Exhibitions and ASME, available: <http://www.reedshows.com/nmw/handouts/2D21.doc>.
- “Education for the Future Workforce,” in *Proceedings, ASEE Mid-Atlantic Section Spring Conference* (co-authored with A. J. Foote and S. C. Pender), Loyola College, MD, April 24–25, 2009.

REFERENCES

1. K. M. Gardiner, “Management Structures for Realization of High Productivity,” in *Proceedings of the International Conference on Productivity and Quality Research (ICPQR-93)*, Vol. 1, Industrial Engineering and Management Press (IIE), Norcross, GA, 1993, pp. 237–246.
2. K. M. Gardiner, “An Integrated Design Strategy for Future Manufacturing Systems,” *J. Manufacturing Syst.*, **15** (1), 52–61, January 1996.
3. M. Wood, *The Story of England*, reprint ed., Penguin, London, 2012; “Michael Wood’s Story of England,” DVD, available: www.pbs.org/programs/michael-woods-story-england/ or http://www.britainexpress.com/History/Feudalism_and_Medieval_life.htm, accessed October, 2012.
4. E. Burton and P. J. Marique, *The Catholic Encyclopedia*, Vol. VII, Robert Appleton Co., 2003, available: <http://www.newadvent.org/cathen/07066c.htm>.

5. G. Agricola, *De Re Metallica*, translated from the first Latin edition of 1556 by H. C. Hoover and L. H. Hoover, Dover Publications, New York, 1950.
6. http://en.wikipedia.org/wiki/Venetian_Arsenal, accessed October 2012.
7. D. Cardwell, *The Fontana History of Technology*, Fontana, London, 1994.
8. A. Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*, Vol. I, Everyman's Library, Dent & Sons, London, 1904.
9. N. Machiavelli, *The Prince*, Bantam Classics, New York, 1984.
10. Sun Tzu, *The Art of War*, Dover Publications, New York, 2002.
11. F. W. Taylor, *The Principles of Scientific Management*, unabridged republication, Dover Publications, Mineola, NY, 1998; original, Harper & Brothers, New York 1919.
12. "Gives \$10,000,000 to 26,000 employees—Minimum Wage \$5 a Day," *New York Times*, January 6, 1914, available: <http://www.nytimes.com/learning/general/onthisday/big/0105.html>, accessed October 2012.
13. U. B. Sinclair, *The Jungle*, uncensored original edition, Sharp Press, Tucson, AZ; new edition, April 1, 2003.
14. D. A. Hounshell, *From the American System to Mass Production 1800–1932 — The Development of Manufacturing Technology in the United States*, Johns Hopkins University Press, Baltimore, MD, 1984.
15. Bureau of Labor Statistics (BLS), Department of Commerce, News Release, January 27, 2012, available: <http://www.bls.gov/news.release/pdf/union2.pdf>, accessed October, 2012.
16. B. G. Hoffman, *New American Icon: Alan Mulally and the Fight to Save Ford Motor Company*, Crown Business, New York, 2012.
17. K. M. Gardiner, T. E. Schlie, and N. L. Webster, "Manufacturing Globalization and Human Factors," in *Proceedings of the Sixteenth International Conference on Flexible Automation and Intelligent Manufacturing*, Limerick University, Ireland, June 2006.
18. "Schumpeter, Joseph (1883-1950); Joseph Schumpeter /Creative Destruction" and *The Concise Encyclopedia of Economics* (© 1999–2008), available: <http://www.econlib.org/library/Enc/bios/Schumpeter.html>, "Can capitalism survive? No. I do not think it can." and http://en.wikipedia.org/wiki/Creative_destruction, accessed October 27, 2012.
19. T. Kidder, *The Soul of a New Machine*, Atlantic-Little, Brown and Company, Boston, 1981.
20. F. Guterl, "Design Case History: 'Apple's MacIntosh,'" *IEEE Spectrum*, **21**(12), 34–44, December 1984.
21. P. F. Drucker, "The Coming of the New Organization," *Harvard Business Rev.*, **66**(1), 45–53, January–February 1988.
22. P. F. Drucker, "The Emerging Theory of Manufacturing," *Harvard Business Rev.*, **68**(3), 94–102, May–June 1990.
23. A. S. Grove, *Only the Paranoid Survive, Currency*, Doubleday, New York, 1996.
24. M. Dell, *Direct from Dell: Strategies that Revolutionized an Industry*, HarperCollins, New York, 2000.
25. R. J. Trent, *Strategic Supply Management: Creating the Next Source of Competitive Advantage*, J. Ross Publishing, New York, 2007.
26. P. F. Drucker, "Knowledge Work and Knowledge Society—The Social Transformations of this Century," Edwin L. Godkin lecture, Harvard University, May 4, 1994.
27. G. Hardin, *Filters Against Folly: How to Survive Despite Ecologists, Economists, and the Merely Eloquent*, Viking Penguin, New York, 1986.
28. http://en.wikipedia.org/wiki/Cloud_computing, <http://computer.howstuffworks.com/cloud-computing/cloud-computing.htm>, accessed October 2012.
29. http://en.wikipedia.org/wiki/Product_lifecycle_management or vendor sites at <http://www-01.ibm.com/software/plm/>.
30. J. R. Evans and W. M. Lindsay, *Managing for Quality and Performance Excellence*, 8th ed., South-Western, Cengage Learning, Mason, OH, 2011.
31. J. R. Evans and W. M. Lindsay, *An Introduction to Six Sigma & Process Improvement*, South-Western, Cengage Learning, Mason, OH, 2005.

32. J. Liker, *The Toyota Way: 14 Management Principles from the World's Greatest Manufacturer*, McGraw-Hill, New York, 2003.
33. L. V. Gerstner, "Who Says Elephants Can't Dance? Inside IBM's Historic Turnaround," HarperBusiness, New York, 2002.
34. "Teamwork—The Team Member Handbook," Pritchett & Associates, available: <http://www.PritchettNet.com>; "The Team Memory Jogger," GOAL/QPC and Oriol Inc., Salem, NH, 1995, available: <http://www.goalqpc.com>.
35. C. Clipson, "Design as a Business Strategy," in *Handbook of Design Management*, Basil Blackwell, Oxford, 1990, pp. 96–105; private communication concerning Saturn Project.
36. *The American Heritage Dictionary*, 2nd college ed., Houghton Mifflin, Boston, 1982.
37. P. Hawken, A. , and L. H. Lovins, *Natural Capitalism: Creating the Next Industrial Revolution*, Little Brown & Company, New York, 2000.
38. L. H. Lovins and B. Cohen, *Climate Capitalism: Capitalism in the Age of Climate Change*, Hill & Wang, New York, 2011.
39. E. O. Wilson, *The Future of Life*, Knopf, New York, 2002.
40. E. Schlosser, *Fast Food Nation—The Dark Side of the all-American Meal*, Perrenial, HarperCollins, New York, 2002.
41. S. L. Goldman, R. N. Nagel, and K. Preiss, *Agile Competitors and Virtual Organizations—Strategies for Enriching the Customer*, Van Nostrand Reinhold, Princeton, NJ, 1994.
42. R. Dove, *Response Ability—The Language, Structure, and Culture of the Agile Enterprise*, Wiley, New York, 2001.
43. Kenneth Preiss, Stephen L. Goldman, and Roger N. Nagel, *Cooperate to Compete: Building Agile Relationships*, Wiley, Hoboken, NJ, 1996.
44. A. S. Brown, "Staying Alive—Forget Competing with China on Price. These U.S. Manufacturers Have Found Ways to Earn Their Bread," *Mech. Eng.*, **128**(1), 22–26, January 2006.
45. J. Dwyer, "A Manufacturing About-Face: Made in America But Sold in China," *New York Times*, September 21, 2012, available: <http://www.nytimes.com/2012/09/21/nyregion/a-manufacturing-about-face-made-in-america-but-sold-in-china.html>, accessed October 27, 2012.
46. P. B. B. Turney, *Common Cents—The ABC of Performance Breakthrough*, Cost Technology, Inc., Beaverton, OR, 1991.
47. J. Stiglitz, "The Benefits of Globalization Are Unevenly Spread," quoted in *The Guardian Weekly*, March 10–16, 2006.
48. J. E. Stiglitz, *The Price of Inequality*, W. W. Norton & Company, New York, 2012.
49. B. K. Hunnicutt, "Kellogg's Six Hour Day," Economic History Association, September 10, 1998, available: http://eh.net/book_reviews/kelloggs-six-hour-day, accessed October 2012.
50. P. H. Diamandis and S. Kottler, *Abundance—The Future Is Better Than You Think*, Free Press, New York, 2012.

CHAPTER 2

ENVIRONMENTALLY BENIGN MANUFACTURING

William E. Biles
University of Louisville
Louisville, Kentucky

1 INTRODUCTION	29	4.3 Metal-Forming Processes	38
2 ENVIRONMENTALLY BENIGN MANUFACTURING	29	4.4 Metal-Joining Processes	41
3 MANUFACTURING AND SUPPLY CHAIN	30	4.5 Plastic Injection Molding	45
3.1 Tier I and Tier II Suppliers	30	5 MANUFACTURED PRODUCT	50
3.2 Transporters	31	REFERENCES	50
4 MANUFACTURING PROCESSES	31	BIBLIOGRAPHY	50
4.1 Machining Processes	31		
4.2 Metal Casting	32		

1 INTRODUCTION

How might mankind enjoy the fruits of an advanced civilization without endangering the viability of planet Earth for future generations? That is the fundamental challenge that we confront in the twenty-first century. In a time when the comforts and pleasures that can be derived from the products of modern technology are accessible for a significant portion of the world's population, how can we manufacture and deliver those products in an environmentally benign fashion?

2 ENVIRONMENTALLY BENIGN MANUFACTURING

The *environmentally benign manufacturing* movement addresses the dilemma of maintaining a progressive worldwide economy without continuing to damage our environment. How can companies—driven by the necessity for manufacturing the products sought by their customers in a cost-effective manner while maintaining market share and providing gains for their stockholders—also heed the growing clamor for a safe environment? This dilemma is fundamentally a trade-off between the needs of current generations and those of future generations. Will we seek creature comforts for ourselves without regard to the safety and well-being of our children and our children's children? Or will we reach a compromise that allows current generations to reap the benefits of our modern technological society while assuring the same benefits for future generations? The challenge for environmentally conscious manufacturers is to find ways to factor both economic and environmental considerations into their business plans.

The fundamental issue in environmentally benign manufacturing is to align business needs with environmental needs. That is, how do we manufacture market-competitive products without harming the air, water, or soil on planet Earth? How do we motivate companies to behave unilaterally to adopt environmentally benign manufacturing practices? Will nation states unilaterally recognize the need to impose environmental standards on companies manufacturing products within their national boundaries? Recent experience informs us that progress is being made on each of these fronts but that we have a long way to go to fully protect the environment from the offenses committed by the worldwide manufacturing community.

3 MANUFACTURING AND SUPPLY CHAIN

The issue of environmentally benign manufacturing is not isolated on the manufacturing function. Environmental issues abound from tier I and II suppliers to the manufacturing system all the way through the supply chain to the consumer. Figure 1 shows the position of the manufacturing function in the overall supply chain.

3.1 Tier I and Tier II Suppliers

Each tier I or II supplier has its own *manufacturing processes*, each with its own environmental impacts. It is incumbent upon the primary manufacturer to qualify its tier I and II suppliers not only in terms of quality, cost, and on-time delivery but on their environmental performance as well. Suppliers must be made to understand that their very financial viability depends on their adopting sound environmental practices. Their role in the supply chain cannot be ignored. It is the responsibility of the primary manufacturer to ensure that its tier I and II suppliers adhere to environmental standards.

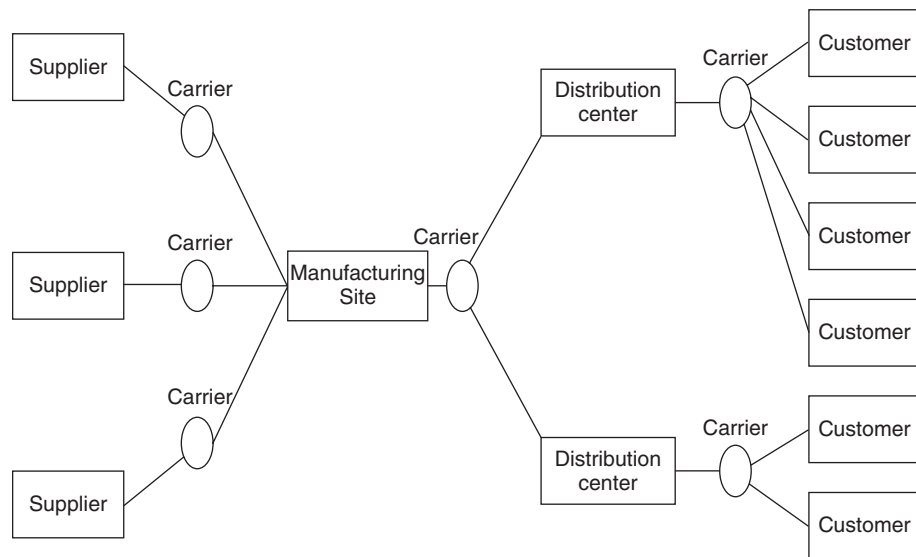


Figure 1 Material and information flow in the supply chain: Material flow is usually left to right, information flow right to left.

3.2 Transporters

The transportation function in the supply chain is also important in terms of its environmental impacts. *Transporters* are those entities that move materials and products from one point to another in the supply chain. Transporters are typically selected and retained according to their cost and reliability performance. Scant attention is paid to the issue of energy expenditure per unit delivery. In an *environmentally conscious manufacturing* approach, primary manufacturers must give closer attention to *energy expenditure per unit delivery* in selecting the mode of transportation from among highway, rail, air, water, and pipeline.

Cost and delivery time considerations must be balanced against energy expenditure in choosing the transportation mode. For example, consider the case of a refrigerator manufactured in the United States which is to be shipped to a distribution facility located 500 miles away. It is probably reasonable to immediately exclude pipeline (infeasible), water (not accessible), and air (too costly) transporters from consideration in this application. The trade-off between highway and rail—both of which are feasible, accessible, and within acceptable cost boundaries for the transport of refrigerators—should incorporate a comparison of the energy expenditure per unit (refrigerator) transported. Such a comparison would very likely come down in favor of rail transportation in terms of both cost and energy expenditure and in favor of highway in terms of delivery time. At present, the delivery time consideration dominates the transporter selection decision in favor of highway transportation. The entire transporter selection issue needs to be reexamined to consider environmental effects of the supply chain transportation function.

4 MANUFACTURING PROCESSES

The manufacturing process itself is perhaps the most important stage in the supply chain in terms of overall environmental impact. Here we shall consider five manufacturing processes that apply to metals and plastics: (1) machining processes, (2) metal casting, (3) metal forming, (4) metal joining, and (5) plastics injection molding.

4.1 Machining Processes

Machining processes include such manufacturing operations as turning, milling, drilling, boring, thread cutting and forming, shaping, planing, slotting, sawing, shearing, and grinding.¹ Each of these processes involves the removal of metal from stock such as a cylindrical billet, cylindrical bar stock, or a cubical block. Metal-cutting economics seek to (1) minimize the cost of the metal-cutting operation, (2) maximize tool life, or (3) maximize production rate. An environmentally benign manufacturing approach would add *minimizing environmental impact* to this list of economic objectives.

The achievement of these economic objectives in machining requires the use of *cutting fluids*, which act as coolants and/or lubricants in the machining process. The four major types of cutting fluids are (1) soluble oil emulsions with water-to-oil ratios ranging from 20:1 to 80:1, (2) oils, (3) chemicals and synthetics, and (4) air. Cutting fluids have six major roles in machining:

1. Removing the heat of friction
2. Minimizing part deformation due to heat
3. Reducing friction among chips, tool, and workpiece
4. Washing away chips

5. Reducing possible corrosion on both the workpiece and machine
6. Preventing built-up edges on the product or part

The environmental impacts of machining processes are principally of two types: (1) the accumulation of metal chips and (2) the release of cutting fluids into the environment. The best solution to the problem of chip accumulation is to recycle them by incorporating them as charge into the metal-casting operation. But recycling may involve transporting the chips to a distant site, thereby incurring the *transporter* impact. The best way to handle cutting fluids is to recycle them back to the machining operation, which requires that chips be separated from the machining effluent and that the cutting fluid be reconstituted to as close to its original state as possible. Each of these steps incurs an economic cost which must be balanced against the cost of the environmental impact of simply placing the chips and used cutting fluid into a waste site.

Electrical discharge machining (EDM) removes electrically conductive material from the raw material stock by means of rapid, repetitive spark discharges from a pulsating dc power supply, with dielectric flowing between the workpiece and the tool (Fig. 2). The cutting tool (electrode) is made of an electrically conductive material, usually carbon. The shaped tool is fed into the workpiece under servocontrol. A spark discharge then breaks down the dielectric fluid. The frequency and energy per spark are set and controlled with a dc power source. The servocontrol maintains a constant gap between the tool and the workpiece while advancing the electrode. The dielectric oil acts as a cutting fluid, cooling and flushing out the vaporized and condensed material while reestablishing insulation in the gap. Material removal rate ranges from 16 to 245 cm³/h. EDM is suitable for cutting materials regardless of their hardness or toughness. Round or irregularly shaped holes 0.002 in. (0.05 mm) in diameter can be produced with *L/D* ratio of 20:1. Narrow slots with widths as small as 0.002–0.010 in. (0.05–0.25 mm) can be cut by EDM.

4.2 Metal Casting

Metal-casting processes are divided according to the specific type of molding method as follows: (1) sand casting; (2) die casting; (3) investment casting; (4) centrifugal casting; (5) plaster mold casting; and (6) permanent casting. This section discusses the first three of these.²

Sand Casting

Sand casting is one of the most ancient forms of metalworking. The first sand casting of copper dates to about 6000 years ago. Sand casting consists of pouring molten metal into shaped

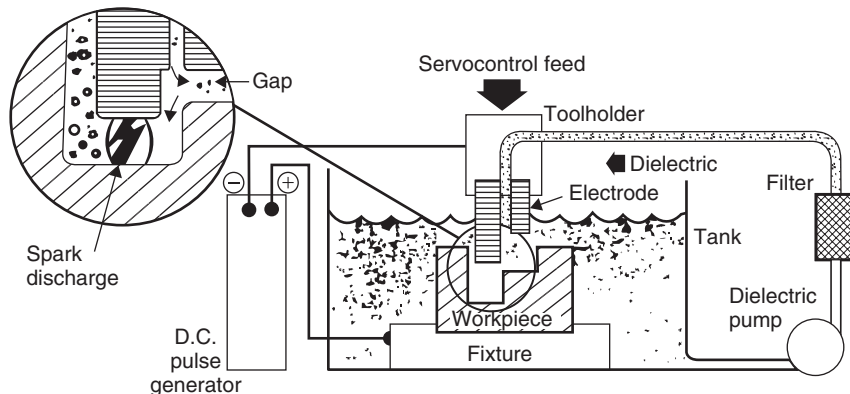


Figure 2 Electrical discharge machining.

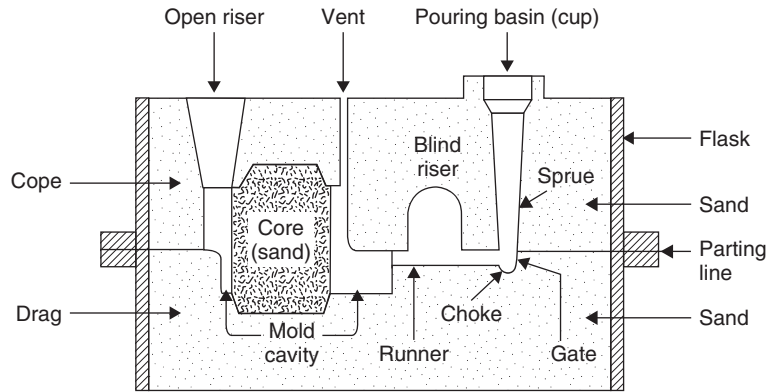


Figure 3 Sectional view of a sand-casting mold.

cavities formed in a sand mold, as shown in Fig. 3. The sand used in fabricating the mold may be natural, synthetic, or artificially blended material.

Sand casting is a relatively simple process and consists of the following steps:

1. Mold preparation
2. Core preparation
3. Core setting
4. Metal preparation
5. Metal pouring
6. Part shakeout
7. Part cleaning
8. Sand reclamation
9. Sprue and gate reclamation

This section describes each step.

Mold Preparation. A mold is fabricated from foundry sand. It is created by pouring and compacting sand around a pattern. Once the sand is compacted, the pattern is withdrawn, leaving a cavity in the shape of the part to be produced. The cavity holds the molten metal in the desired shape until it cools. Molding *sand* is a mixture of approximately 85% sand, from 4 to 10% clay, and from 2 to 5% water by mass. Small quantities of additives are used to prevent the metal from oxidizing as it cools. These additives are usually bituminous coal, anthracite, or ground coke.

Core Preparation. Cores are necessary for parts that are especially complicated or have internal cavities. Cores are created from sand and a binder—usually in the form of a resin—that cures through heat or gasification. The sand and binder are put in a mold called a *core box* that forms the desired shape. They are then removed from the core box and allowed to cure before placing them in the sand casting mold. The placement of the core is illustrated in the left half of Fig. 3.

Core Setting. Once the mold and cores have been prepared, the cores are set in place inside the mold and the mold is closed.

Metal Preparation. Metal—usually iron, steel, or aluminum—is prepared by melting ingots or scrap with the additives or alloying materials needed to give the finished product its desired properties. Most sand casting is accomplished by melting and blending scrap material.

Metal Pouring. Metal is poured manually from a ladle or tilting furnace, or most commonly from an automatic pouring ladle, that is charged from holding furnaces.

Shakeout. Molds containing cooled parts are transferred by conveyor to a large rotary drum, where the sand molds are broken and the sand is separated from the newly molded parts.

Part Cleaning. Usable parts are separated from gates and risers, and damaged or incompletely formed parts are sorted out. Further cleaning may also be accomplished in the form of pressing, hand grinding, sandblasting, or tumbling the parts to remove the parting lines and rough edges as well as any burnt sand.

Sand Reclamation. All modern foundries reclaim molding sand for reuse. The sand is run through a process where lumps are broken up and any solids are removed by screening. New sand, clay, and water are added as needed to return the sand to a usable condition. Some sand that cannot be reclaimed is discarded. Most foundries have a sand laboratory whose responsibility is to monitor and manipulate the condition of the molding sand.

Sprue and Gate Reclamation. Any metal that is not a usable part is returned to the scrap area to be used in a future melt.

Environmental Concerns with Sand Casting

With respect to sand casting, the environmentally benign manufacturing function is concerned with minimizing the impact of the manufacturing steps just listed on the environment by changing or replacing processes that produce an environmentally offensive result or hazard. Consideration will be given here to each of the sand-casting subprocesses.

Molds and Cores. Molds and cores are made from sand. For every ton of castings produced, the process requires about 5.5 tons of sand. Problems occur when the sand and binders are exposed to the heat of the molten metal and sometimes during curing processes of mold preparation. This releases a wide variety of organic pollutants that are regulated by the *Clean Air Act* or the *Clean Water Act*. These pollutants come primarily from the chemical binders use to make cores stronger or in some cases from binders added to the sand. When stronger molds are required, chemical binders are added to the mold sand. These binders include furanes, phenolic urethanes, and phenolic esters. The binder is chosen depending on the strength required for the metal being cast and the size of the mold. Other concerns are molding sand additives used to prevent the metal from oxidizing as it cools. These additives are usually bituminous coal, anthracite, or ground coke. Although these additives are a very small component by mass, as they burn off on contact with the molten metal they create an assortment of hazardous air pollutants. Table 1 shows several pollutants associated with binders used in mold preparation.

Metal Preparation and Pouring. While the use of scrap metal can contribute to pollutants, the most significant contributors for these subprocesses are related to the heat input to melt the metal. Pollutants include large amounts of particulates and carbon monoxide as well as smaller amounts of SO₂ and volatile organic carbon (VOC). Emissions are dependent on the type of furnace being used. Electric furnaces have a reduced environmental impact as compared to coke-fired furnaces of older foundries. Many foundries use pollution control technology in the form of scrubbers to clean air before releasing it to the outside. These are used on all types of

Table 1 Selected Pollutants Associated with Binders Used in Mold Preparation

	Benzene	Methanol	Phenol	Toluene	Formaldehyde	MMDI
Furane	•	•	•	•		
Phenol urethane			•		•	•
Phenol ester			•		•	

Note: MMDI is an acronym for monomeric methylene diphenyl diisocyanate.

Table 2 Approximate On-Site Emissions from Various Furnaces (lb./ton Metal)

	PM	CO	So ₂	VOC
Fuel-fired reverberatory furnace	2.2	Unknown	N/A	Unknown
Induction furnace	1	~0	~0	Unknown
Electric arc furnace	12.6	1–38	~0	0.06–0.30
Coke-fired cupola	13.8	146	1.25+	Unknown

Note: Does not include emissions from electricity generation or fuel extraction.

Table 3 Energy Requirements at Foundry: Saleable Cast Material for Foundry Furnaces

Fuel Source	Furnace Type	MBtu/ton
Fuel-fired	Crucible	1.8–6.8
	Reverberatory	2.5–5.0
	Cupola (coke)	5.8
	Cupola (NG)	1.6
Electric	Induction	4.3–4.8
	Electric arc	4.3–5.2
	Reverberatory	5.2–7.9
	Cupola	1.1

furnaces. Wet scrubbers are also used but are less common and are used primarily on coke-fired cupola furnaces. These methods are effective at controlling air emissions, but they produce waste streams in the form of solid waste or contaminated water, which must be processed further. Table 2 shows pollutants generated by melting metal for several types of furnaces. Table 3 gives the energy requirements at the foundry for both fuel-fired and electric furnaces.

Cleaning. Cleaning the product can involve the use of organic solvents, abrasives, pressurized water, or acids, often followed by protective coatings. Techniques used to remove sand and flashing include vibrating, wire brushing, blast cleaning band saws, cutoff wheels, and grinders.

Removing Sprues, Runners, and Flashing. Although particles and effluent pollutants are created in this stage, they are largely contained by filters and closed systems.

Sand Reclamation. Up to 90% of molding sand can be reused in a green sand foundry after filtration for fine dust and metal particles. Sand with chemical binders can be used only in small quantities, however. Sand that is not reused is sometimes used in road bases and asphalt concrete.³ In the United States, from 7 to 8 million tons of mold sand (about 0.5 tons of sand/ton of cast metal) per year ends up in landfills. Spent sand makes up almost 70% of foundry solid wastes.⁴

From an environmental perspective, the foundry industry has improved remarkably in recent years. U.S. and off-shore foundries have been forced by both legislation and automakers to reduce their pollutants and waste streams—hence, the positive influence of manufacturers (automakers) on tier I suppliers (foundries). Foundries are relying more on electric and natural gas furnaces, thereby reducing the amount of input energy required and minimizing the amount of pollutants. Sand reclamation and use of spent sand for other purposes reduces the impact on landfills. The recent use of trimming presses helps to eliminate the need to grind parts to remove gates and sprues. One of the areas that could benefit from continued research is the development of benign binders for core and mold making processes. Redesigning parts to eliminate cores would also be helpful.

Another environmental concern for the sand-casting process is the generation of waste from the machining of cast metal parts. Machining allowances are required in many cases because of unavoidable surface impurities, warpage, and surface variations. Average machining allowances are given in Table 4. Good practice dictates use of the minimum section thickness compatible with the design. The normal section recommended for various metals is shown in Table 5 (see Ref. 1).

Table 4 Machining Allowances for Sand Castings (in./ft.)

Metal	Casting Size (in.)	Finish Allowance
Cast irons	Up to 12	3/32
	13–24	1/8
	25–42	3/16
	43–60	1/4
	61–80	5/16
	81–120	3/8
Cast steels	Up to 12.	1/8
	13–24.	3/16
	25–42	5/16
	43–60	3/8
	61–80	7/16
	81–120	1/2
Malleable irons	Up to 8	1/16
	9–12	3/32
	13–24	1/8
	25–36	3/16
Nonferrous metals	Up to 12	1/16
	13–24	1/8
	25–36	5/32

Table 5 Minimum Sections for Sand Castings (in./ft.)

Metal	Section
Aluminum alloys	3/16
Copper alloys	3/32
Gray irons	1/8
Magnesium alloys	5/32
Malleable irons	1/8
Steels	1/4
White irons	1/8

Die Casting

Die casting may be classified as a permanent-mold-casting system. However, it differs from the process just described in that molten metal is forced into the mold or die under high pressure [1000–30,000 psi (6.89–206.8 MPa)]. The metal solidifies rapidly (within a fraction of a second) because the die is water cooled. Upon solidification, the die is opened. Ejector pins automatically eject the casting from the die. If the parts are small, several of them may be made at one time in what is termed a *multicavity die*.

There are two main types of machines used: the hot-chamber and the cold-chamber types.

Hot-Chamber Die Casting. In the hot-chamber machine, the metal is kept in a heated holding pot. As the plunger descends, the required amount of alloy is automatically forced into the die. As the piston retracts, the cylinder is again filled with the right amount of molten metal. Metals such as aluminum, magnesium, and copper tend to alloy with the steel plunger and cannot be used in the hot chamber.

Cold-Chamber Die Casting. This process gets its name from the fact that the metal is ladled into the cold chamber for each shot. This procedure is necessary to keep the molten-metal contact time with the steel cylinder to a minimum. Iron pickup is prevented, as is freezing of the plunger in the cylinder.

Advantages and Limitations. Die-casting machines can produce large quantities of parts with close tolerances and smooth surfaces. The size is limited only by the capacity of the machine. Most die castings are limited to about 75 lb (34 kg) of zinc; 65 lb (30 kg) of aluminum; and 44 lb (20 kg) of magnesium. Die casting can provide thinner sections than any other casting process. Wall thicknesses as thin as 0.015 in. (0.38 mm) can be achieved with aluminum in small items. However, a more common range on larger sizes will be 0.105–0.180 in.

Some difficulty is experienced in getting sound castings in the larger capacities. Gases tend to be entrapped, which results in low strength and annoying leaks, causing an air pollution problem. One way to reduce metal sections without sacrificing strength is to add ribs and bosses into the product design. An approach to the porosity problem has been to operate the machine under vacuum.

The surface quality of the casting is dependent on that of the mold. Parts made from new or repolished dies may have a surface roughness of 24 $\mu\text{in.}$ (0.61 μm). A high surface finish means that, in most cases, coatings such as chrome plating, anodizing, and painting may be applied directly. More recently, decorative texture finishes are obtained by photoetching. This technique has been used to simulate wood grain finishes as well as textile and leather finishes and to obtain checkering and crosshatching patterns in the surface finish.

Investment Casting. Casting processes in which the pattern is used only once are variously referred to as *lost-wax* or *precision-casting* processes. These involve making a pattern of the desired form out of wax or plastic (usually polystyrene). The expendable pattern may be made by pressing the wax into a split mold or by using an injection-molding machine. The patterns may be gated together so that several parts can be made at once. A metal flask is placed around the assembled patterns, and a refractory mold slurry is poured in to support the patterns and form the cavities. A vibrating table equipped with a vacuum pump is used to eliminate all the air from the mold. Formerly, the standard procedure was to dip the patterns in the slurry several times until a coat was built up. This is called the *investment process*. After the mold material has set and dried, the pattern material is melted and allowed to run out of the mold.

The completed flasks are heated slowly to dry the mold and to melt out the wax, plastic, or whatever pattern material was used. When the molds have reached a temperature of 100°F

(37.8°C), they are ready for pouring. Vacuum may be applied to the flasks to ensure complete filling of the mold cavities. When the metal has cooled, the investment material is removed by vibrating hammers or by tumbling. As with other castings, the gates and risers are cut off and ground down.

Ceramic Process. The ceramic process is somewhat similar to the investment casting in that a creamy, ceramic slurry is poured over a pattern. In this case, however, the pattern, made out of plastic, plaster, wood, metal, or rubber, is reusable. The slurry hardens on the pattern almost immediately and becomes a strong green ceramic of the consistency of vulcanized rubber. It is lifted off the pattern, while it is still in the rubberlike phase. The mold is ignited with a torch to burn off the volatile portion of the mix. It is then put in a furnace and baked at 1800°F (982°C), resulting in a rigid refractory mold. The mold can be poured while still hot.

Full-Mold Casting. Full-mold casting may be considered a cross between conventional sand casting and the investment technique of using lost wax. In this case, instead of a conventional pattern of wood, metals, or plaster, a polystyrene foam or Styrofoam is used. The pattern is left in the mold and is vaporized by the molten metal as it rises in the mold during pouring. Before molding, the pattern is usually coated with a zirconite wash in an alcohol vehicle. The wash produces a relatively tough skin separating the metal from the sand during pouring and cooling. Conventional boundary sand is used in backing up the mold.

4.3 Metal-Forming Processes

Metal-forming processes use a remarkable property of metals—their ability to flow plastically in the solid state without concurrent deterioration of properties. Moreover, by simply moving the metal to the desired shape, there is little or no waste. Figure 4 shows some of the metal-forming processes. Metal-forming processes are classified into two categories: hot-working processes and cold-working processes.

Hot Working

Hot working is defined as the plastic deformation of metals above their recrystallization temperature. Here it is important to note that the crystallization temperature varies greatly with different materials. Lead and tin are hot worked at room temperature, while steels require temperatures of 2000°F (1100°C). Thus, hot working does not necessarily imply high absolute temperatures.

Hot working can produce the following improvements in metal products:

1. Grain structure is randomly oriented and spherically shaped, which results in a net increase not only in the strength but also in ductility and toughness.
2. Inclusions or impurity material in metal are reoriented. The impurity material often distorts and flows along with the metal.
3. This material, however, does not recrystallize with the base metal and often produces a fiber structure. Such a structure clearly has directional properties, being stronger in one direction than in another. Moreover, an impurity originally oriented so as to aid crack movement through the metal is often reoriented into a “crack arrestor” configuration perpendicular to crack propagation.

Isothermal Rolling

The ordinary rolling of some high-strength metals, such as titanium and stainless steels, particularly in thicknesses below about 0.15 in. (3.8 mm), is difficult because the heat in the sheet is

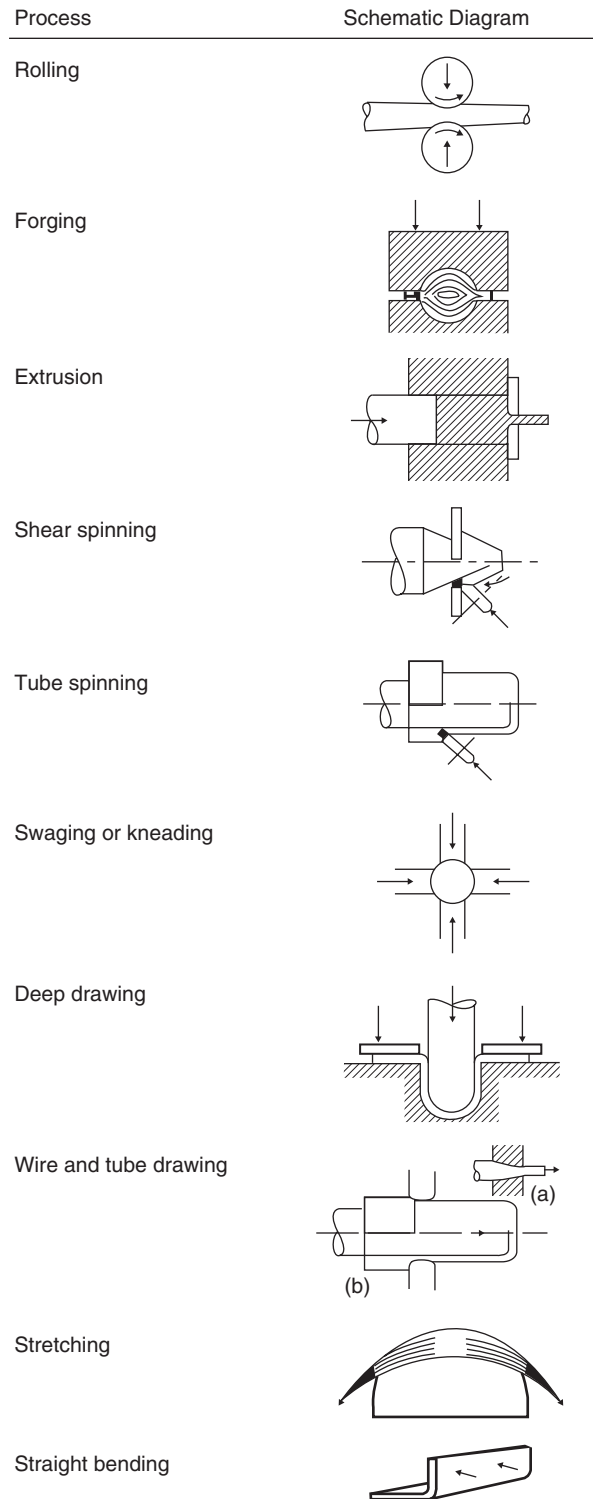


Figure 4 Several metal-forming processes.

transferred rapidly to the cold and much more massive rolls. This difficulty has been overcome by *isothermal rolling*. Localized heating is accomplished in the area of deformation by the passage of a large electrical current between the rolls through the sheet. Reductions up to 90% per roll have been achieved. The process usually is restricted to widths below 2 in. (50 mm).

Forging

Forging is the plastic working of metal by means of localized compressive forces exerted by manual or power hammers, presses, or special forging machines. Various types of forging have been developed to provide great flexibility, making it economically possible to forge a single piece or to mass produce thousands of identical parts. The metal may be drawn out, increasing its length and decreasing its cross section; upset, increasing the cross section and decreasing the length; or squeezed in closed impression dies to produce multidirectional flow. The state of stress in the work is primarily uniaxial or multiaxial compression. The most common forging processes are as follows:

- Open-die hammer
- Impression die drop forging
- Press forging
- Upset forging
- Roll forging
- Swaging

Extrusion

In the extrusion process shown in Fig. 5, metal is compressively forced to flow through a suitably shaped die to form a product with a reduced cross section. Although extrusion may be performed either hot or cold, hot extrusion is employed for many metals to reduce the forces required, to eliminate cold-working effects, and to reduce directional properties. The stress state within the material is triaxial compression.

Lead, copper, aluminum, and magnesium and alloys of these metals are commonly extruded, taking advantage of the relatively low yield strengths and extrusion temperatures. Steel is more difficult to extrude. Yield strengths are high and the metal has a tendency to weld to the walls of the die and confining chamber under the conditions of high temperatures and pressures. With the development and use of phosphate-based and molten glass lubricants, however, substantial quantities of hot steel extrusions are now produced. These lubricants adhere to the billet and prevent metal-to-metal contact throughout the process.

Almost any cross-sectional shape can be extruded from the nonferrous metals. Hollow shapes can be extruded by several methods. For tubular products, the stationary or moving mandrel process is often employed. For more complex internal cavities, a spider mandrel or torpedo die is used. Obviously, the cost for hollow extrusions is considerably greater than for solid ones, but a wide variety of shapes can be produced that cannot be made by any other process.

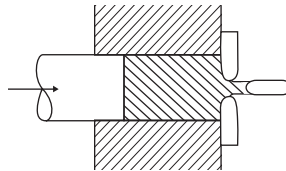


Figure 5 Metals extrusion process.

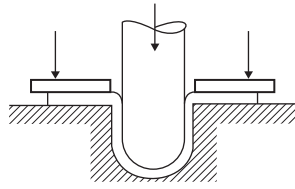


Figure 6 Deep drawing of metal part.

Drawing

Drawing, shown in Fig. 6, is a process for forming sheet metal between an edge-opposing punch and a die (draw ring) to produce a cup, cone, box, or shell-like part. The work metal is bent over and wrapped around the punch nose. At the same time, the outer portions of the blank move rapidly toward the center of the blank until they flow over the die radius as the blank is drawn into the die cavity by the punch. The radial movement of the metal increases the blank thickness as the metal moves toward the die radius; as the metal flows over the die radius, this thickness decreases because of the tension in the shell wall between the punch nose and the die radius and (in some instances) because of the clearance between the punch and the die.

4.4 Metal-Joining Processes

The most common forms of metal joining are welding, soldering, and brazing. Each of these processes has the potential to be environmentally offensive by generating noxious gases as part of the joining process or by producing metal wastes that must be disposed of. Degarmo, Black, Kohser, and Klamecki provide an excellent discussion of these various joining processes (and, indeed, any of the manufacturing processes discussed in the chapter).⁵ Figure 7 gives the various classifications of welding processes employed in manufacturing.

Welding is the most common metal-joining process. The principal classes of welding processes include (1) gas-flame welding, which utilizes a high-temperature gas to melt selected surfaces of the mating parts; (2) arc-welding processes, which utilize an electric arc to produce molten material between mating parts; and (3) resistance-welding processes, which utilize both heat and pressure to induce coalescence. Brazing and soldering are utilized when the mating surfaces cannot sustain the high temperatures required for welding mating parts. The ensuing sections give brief discussions of each of these joining processes and describe how environmental offenses can be avoided.

Welding Processes

As just stated, three of the most common classes of welding processes used in manufacturing are oxyfuel gas welding, arc welding, and resistance welding. The coalescence between two metals requires sufficient proximity and activity between the atoms of the pieces being joined to cause the formation of common crystals.

Gas-Flame Processes. Oxyfuel gas-welding processes utilize as their heat source the flame produced by the combustion of a fuel gas and oxygen. The combustion of *acetylene* (C_2H_2)—commonly known as the *oxyacetylene torch*—produces temperatures as high as $5850^\circ F$ ($3250^\circ C$). Three types of flames can be obtained by varying the oxygen–acetylene ratio: (1) If the ratio is between 1:1 and 1.15:1, all oxygen–acetylene reactions are carried to completion and a *neutral flame* is produced; (2) if the ratio is closer to 1.5:1, an *oxidizing flame* is produced, which is hotter than the neutral flame but similar in appearance; and (3) excess fuel produces a *carburizing flame*.

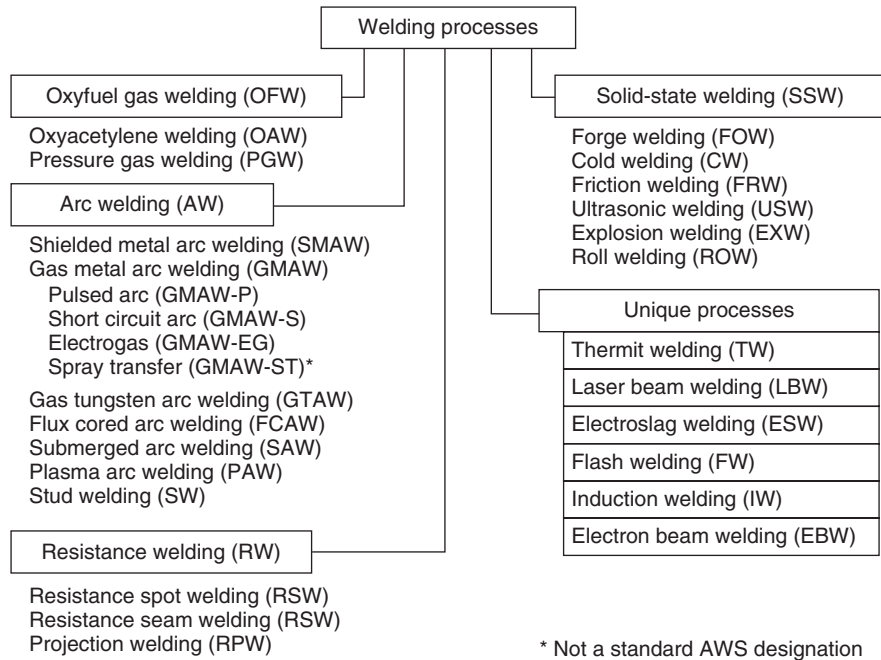


Figure 7 Classification of several common welding processes.

Almost all oxyfuel gas welding is of the *fusion* type, which means that the metals to be joined are simply melted at the interfacing surfaces and no pressure is required. This process is best suited to steels and other ferrous metals. There is a low heat input to the part, and penetrations are only about 3 mm.

The environmental impacts of oxyfuel gas welding include the generation of combustion products, which have to be *scrubbed* before release to the atmosphere, and the production of slag and waste metal that must be safely disposed of.

Arc-Welding Processes. Arc-welding processes employ the basic circuit shown in Fig. 8. Welding currents typically vary from 100 to 1000 A, with voltages in the range of 20–50 V.

In one type of arc-welding process, the electrode is consumed and thus supplies the molten metal. A second process utilizes a nonconsumable tungsten electrode, which requires a separate

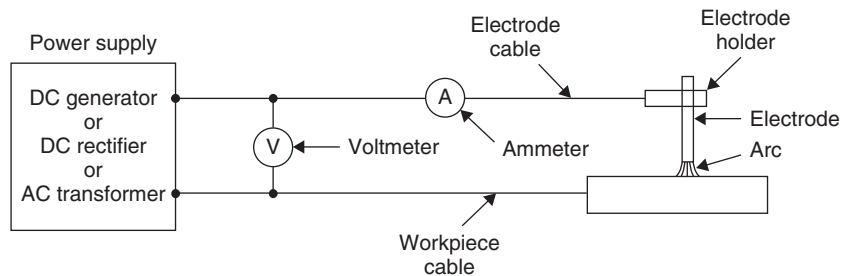


Figure 8 Basic circuit for the arc-welding process.

metal wire to supply the molten metal. Filler materials must be selected to be compatible with the mating surfaces being welded. In applications where a close fit is required between mating parts, gas-tungsten arc welding can produce high-quality, nearly invisible welds.

In *plasma arc welding* an arc is maintained between a nonconsumable electrode and the workpiece in such a way as to force the arc to be contained within a small-diameter nozzle, with an inert gas forced through the stricture. Plasma arc welding is characterized by extremely high (30,000°F) temperatures, which offers very high welding speeds and hence high production rates.

The environmental impacts of arc-welding processes include the generation of metal waste and the requirement for relatively high power.

Resistance-Welding Processes. In *resistance welding*, both heat and mechanical pressure are used to induce coalescence. Electrodes are placed in contact with the material, and electrical resistance heating is utilized to raise the temperatures of the workpieces and the space between them. These same electrodes also supply the mechanical pressure that holds the workpieces in contact. When the desired temperature has been achieved, the pressure exerted by the electrode is increased to induce coalescence. Figure 9 illustrates a typical resistance-welding circuit. It is important to note that the workpieces actually form part of the electrical circuit and that the total resistance between the electrodes consists of three distinct components: (1) the resistance of the workpieces; (2) the contact resistance between the electrodes and the workpieces; and (3) the resistance between the surfaces to be joined.

The most important environmental consideration in resistance welding is the electrical power consumed.

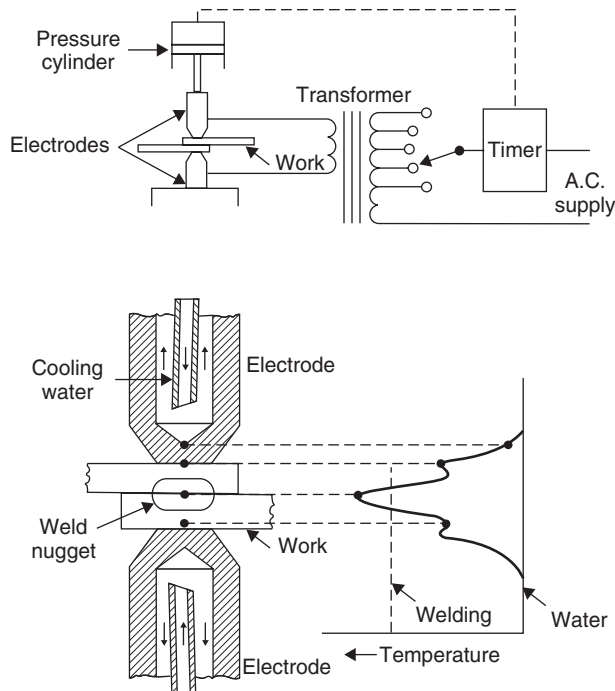


Figure 9 A typical resistance-welding circuit and configuration.

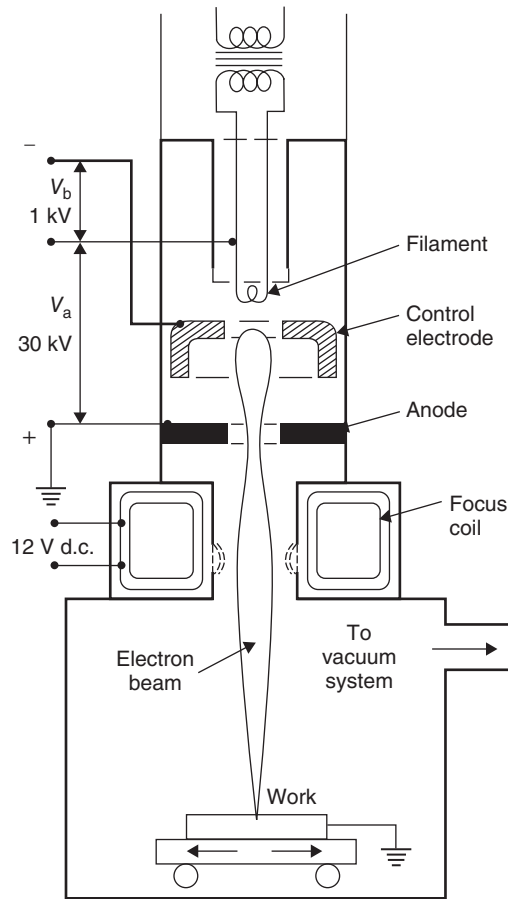


Figure 10 Schematic diagram of electron beam welding process.

Electron Beam Welding. *Electron beam welding* is a fusion welding process which utilizes the heating resulting from the impingement of a beam of high-velocity electrons on the metal parts to be welded. The electron optical system for the electron beam welding process is shown in Fig. 10. An electrical current heats a tungsten filament to about 4000°F, causing it to emit a stream of electrons by *thermal emission*. Focusing coils are employed to concentrate the electrons into a beam, accelerate them, and direct them to a focused spot that is between 0.8 and 3.2 mm in diameter. Since the electrons, which are accelerated at 150 kV, achieve velocities near two-thirds the speed of light, intense heat is generated. Since the beam is composed of charged particles, it can be positioned by electromagnetic lenses. To be effective as a welding heat source, the electron beam must be generated and focused in a high vacuum, typically at pressures as low as 0.01 Pa.

Almost any metal can be welded by the electron beam process, including those that are very difficult to weld by any other process, including tungsten, zirconium, and beryllium. Heat-sensitive metals can be welded without damage to the base metal.

From an environmental standpoint, the absence of shielding gases, fluxes, or filler materials means that the waste material produced by the process is negligible. Only the high-power requirements stand as a problem.

Brazing and Soldering

Brazing is the permanent joining of similar or dissimilar metals through the application of heat and a filler material. Filler metals melt at temperatures as low as 800°F, typically much lower than those of the base metals, which makes brazing a useful joining process for dissimilar metals (ferrous to nonferrous metals, metals with different melting points, or even metal to ceramic). Strong permanent joints are formed by brazing.

Soldering is a type of brazing operation in which the filler material has a melting temperature below 850°F. It is typically used for connecting thin metal pieces, connecting electronic components, joining metals while avoiding high temperatures, and filling surface flaws and defects in metal parts. Soldering can be used to join a wide variety of shapes, sizes, and thicknesses and is widely used to provide electrical coupling or airtight seals. The primary means of heating the filler material is to apply an electrically heated iron rod to melt the filler metal and position it in the proper location on the workpiece. Soldering filler materials are typically low-melting-temperature metals such as lead, tin, bismuth, indium, cadmium, silver, gold, and germanium. Because of their low cost and favorable properties, alloys of tin and lead are most commonly used.

The environmental impacts of brazing and soldering trace to the filler materials used in their application. Since 1988, the use of lead and lead alloys in drinking water lines has been prohibited in the United States. Japan and the European Union prohibit the use of lead in electronic applications.

4.5 Plastic Injection Molding

The injection-molding process involves the rapid pressure filling of a shape-specific mold cavity with a fluid plastic material, followed by the solidification of the material into a product. The process is used for thermoplastics, thermosetting resins, and rubbers.

Principle of Injection Molding

The injection molding of thermoplastics can be subdivided into several stages as illustrated in Fig. 11. At the plastication stage P, the feed unit F operates in much the same way as an extruder,

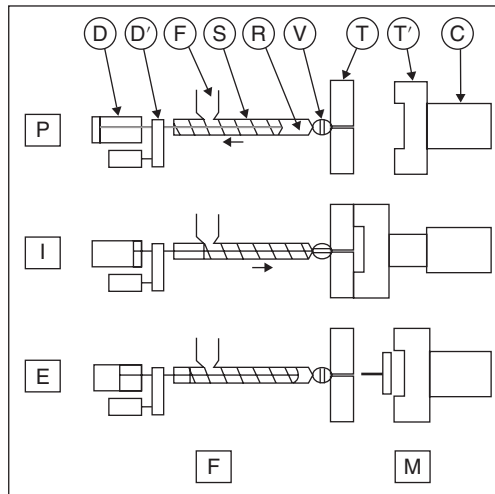


Figure 11 Principle of injection molding. (Source: From Ref. 6, Reprinted with permission of Hanser Publishers, Munich, 1990.)

melting and homogenizing the material in the screw/barrel system. The screw, however, is allowed to retract to make room for the molten material in a space at the cylinder head, referred to here as the material *reservoir*, between the screw tip and a closed valve or an obstruction of solidified material from the previous shot. At the injection stage I, the screw is used as a ram (piston) for the rapid transfer of the molten material from the reservoir to the cavity between the two halves (T and T') of the closed mold. Since the mold is kept at a temperature below the solidification temperature of the material, it is essential to inject the molten material rapidly to ensure complete filling of the cavity. A high holding or packing pressure (10,000–30,000 psi) is normally exerted to partially compensate for the thermal contraction (shrinkage) of the material upon cooling. The cooling of the material in the mold is often the limiting time factor in injection molding because of the low thermal conductivity of polymers. After the cooling stage, the mold can be opened and the solid product removed.

Equipment

Injection-molding machines are now most commonly of the reciprocating screw type, as illustrated in Fig. 12. Two distinct units, referred to as the feed unit F and the mold unit M, are mounted on a frame (F). The feed unit F consists of the plastication/injection cylinder (screw, barrel, and feed hopper), the axial screw drive, and the rotation screw drive.

Although injection-molding machines may occasionally be dedicated to the molding of a single product, a machine is normally used with a variety of tools (molds), which may imply frequent mold changes and the associated costly set-up period. Injection-molding machines are available in a broad range of sizes. They are normally rated by their maximum clamping force, with normal ranges of about 25–150 tons for “small” machines, 150–70 tons for “medium-sized” machines, and 750–5000 tons for “large” machines; the current maximum is 10,000 tons.

Tooling

The interchangeable injection-molding tool, the *mold*, must (1) provide a cavity corresponding to the geometry of the product and (2) allow the ejection of the product after its solidification. Primary mold opening is achieved by fastening one-half of the mold to the stationary platen (T), as shown in Figure 12, and the other half to the moving platen (T'). The stationary mold half is sometimes referred to as the *front, cavity, or negative block* and the moving mold half as the *rear, force, or positive block*. The removal of a product from a cavity surface requires, in addition to an ejection system, a suitable surface finish and an appropriate taper or draft. It need not require a mold release agent.

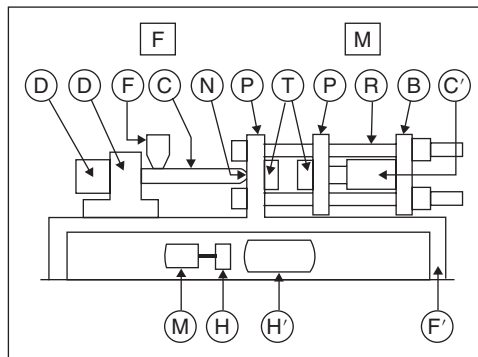


Figure 12 Injection-molding machine. (Source: From Ref. 6, Reprinted with permission of Hanser Publishers, Munich, 1990.)

During injection, the material flows from the nozzle at the tip of the injection unit to the single cavity, or to each of several cavities, through what is referred to here as the *feed system*, generally comprising sprues, runners, and gates. In most cases, injection-molded products need to be removed from one mold half by an ejection (knockout, stripping) device. This device is normally incorporated in the moving mold half. Retractable secondary mold sections may be required when products feature undercuts, reentrant shapes, internal or external threads, and so on.

Runners are machined in mold halves, next to the parting surface. One solution, applicable to chemically stable thermoplastics, consists of having large runners cooled in such a way that a sleeve of insulating solid plastic forms around a molten core, where the intermittent injection flow takes place; this method is referred to as *insulated* or *Canadian* runner molding. Another solution, referred to as *hot runner molding*, involves a heated runner, or manifold block, and is often used in conjunction with valve gating. *Gates* serve several purposes in injection molding. Their easily altered, smaller cross section permits a convenient control of the flow of the molten material, the rapid freezing of the material to shut off the cavity after injection, and the easy separation of the products from the feed appendage (degating). Important savings can be made by using hot runners.

The maximum pressure in injection molds is normally in the range of 4000–12,000 psi corresponding to a clamping force per unit projected area of cavity and feed system in the range of 2–6 tons/in.². The construction of injection molds requires materials with a combination of good thermal conductivity and resistance to mechanical wear and abrasion. Prototype molds can be cast from low-melting alloys. For short-run molds (about 10,000–100,000 moldings), tool steel is normally used. Long runs involving millions of moldings require special hardened and chrome-plated steels.

A variety of techniques are used to form mold cavities: machining of a solid block, computer-aided machining (CAM) centers, hobbing (cold forming), electrochemical machining (ECM), electrical discharge machining (EDM), or spark erosion, electroforming, plating, and etching. For short runs—fewer than 10,000 parts—a mold cavity can be fabricated using a selective-laser-sintering rapid-prototyping process to build a copper-infiltrated iron part.

Auxiliaries

Many thermoplastic resins require thorough drying prior to molding to avoid the formation of voids or a degradation of the material at molding temperatures. Mold temperature control is often achieved by the circulation of a fluid through a separate heater/chiller device. With increased interest in automation, robots have been introduced for the removal of products and feed appendages from open molds and for separation (degating) and sorting. Feed appendages, startup scrap, and occasional production scrap are normally reground in granulators and recycled as a fraction of the feed material.

Materials

All thermoplastics are, in principle, suitable for injection molding, but since fast flow rates are needed, grades with good fluidity (high melt index) are normally preferable.

Products

A major advantage of injection-molded products is the incorporation of fine details such as bosses, locating pins, mounting holes, bushings, ribs, flanges, and so on, which normally eliminates assembly and finishing operations. Thermosetting resin systems such as phenolics (PFs) and unsaturated polyester (UP), often used with fillers or reinforcements, are increasingly injection molded at relatively high speeds. Curing, which involves chemical reactions, takes generally much longer than the injection, and multimold machines are thus often used with

shuttle or rotary systems. Injection molding is increasingly used for producing relatively small rubber products significantly faster than by compression molding and, normally, with a smaller amount of scrap and a better dimensional accuracy. As in the case of thermosetting resins, a heated mold is needed for vulcanization (curing).

Environmental Analysis of Injection-Molding Processes

Plastic components are major parts in electrical and electronic (E&E) products. About 8.5% of the plastic parts produced are for these products. A large number of plastic parts are used in the automobile industry. Some 33% of all small house appliances incorporate plastic components, and about 42% of all plastic materials are used in the manufacture of toys.

This environmental analysis of injection molding highlights a few important points. The type of injection-molding machine (hydraulic, hybrid, or electric) has a large impact on energy consumption. Table 6 shows the energy-related emissions for the injection-molding process, including the compounder stage. Table 7 gives the total annual production of injection-molded plastics. Table 8 gives the total annual energy consumption associated with the production of injection-molded plastics. The impact of injection molding on the environment may seem benign, but it can be significant. We must take into consideration energy consumption, the manufacturing process, and raw material usage. The product life cycle is important because it affects the production, energy, and raw material. The majority of plastic parts that are used in electrical and electronic products are parts made through the injection-molding process. Injection molding involves melting polymer resin together with additives and then injecting the melt

Table 6 Energy-Related Air Emissions for Compounder Stage and Injection Molder Stage

Stage	SEC (MJ/kg)	Energy-Related Emissions				
		CO ₂ (g)	SO ₂ (g)	NO _x (g)	CH ₄ (g)	Hg (mg)
Compounder	5.51	284.25	1.26	0.51	10.32	0.01
Injection molder Hydraulic	13.08	674.82	2.98	1.22	24.29	0.01
hybrid	7.35	379.33	1.68	0.68	13.77	0.01
All electric	6.68	344.57	1.52	0.62	12.50	0.01

Table 7 Injection-Molded Polymer Totals

	Injection Molded ($\times 10^6$ kg/yr)	
	U.S. Only	Global
Six main thermoplastics	5,571	23,899
All plastics	12,031	38,961

Note: The subdivision "6 main thermoplastics" refers to HDPE, LDPE, LLDPE, PP, PS, and PVC.

Table 8 Total Energy Used in Injection Molding (GJ/yr)

Compounder and Injection Molder	U.S.	Global
Six main thermoplastics	9.34×10^7	4.01×10^8
All plastics	2.06×10^8	6.68×10^8

Note: The subdivision "6 main thermoplastics" refers to HDPE, LDPE, LLDPE, PP, PS and PVC.

into the mold to make the final products. This process may have an impact on the environment, but we have to reduce the effect of this process and make it benign as much as possible.

Injection molding is used primarily to produce plastic parts with specific geometrics. The process starts by mixing polymer resin with additives that are specific to the part to gain desired properties such as increased strength. The mix of polymer resin and additives is also combined with colorants if needed at this point and is stored in a hopper. The material is gravity fed into a feeding tube that has screws to push material forward. When in the screws, the material is melted and mixed. The material is fed into the die that will shape the material to the desired part. During the fill stage, hydraulic clamps hold the two ends of the die together until all the necessary material has entered the die and cooled to the desired temperature for removal. Then the clamps release the part and it is removed from the die.

Life Cycle of a Plastic Product

When tracing the life cycle of the process to the beginning, we need to look at how the polymer pellets are manufactured. In injection molding the overall process starts at production stage. This stage takes raw materials from the earth and transforms them with addition of energy into polymers. The raw polymer is shipped in bulk to the compounder, which mixes it with additives in order to give it required properties for application. The polymer is shipped to the injection molder, which transforms the polymer into finished products. The injection molder might add some additives in the process, such as coloring. After being injection molded and packaged, the product is ready for the consumer. When trying to develop the polymer resin used in the injection mold process, the manufacturer uses large amounts of petroleum, and large energy costs are associated with the production of the material. The additives added to the polymer base can be hazardous in large concentrations. The majority of the by-products to the process can be hazardous and are not biodegradable.

Environmental Impact of Plastics Injection Molding. When considering the life cycle of a plastic-based product, it is important to understand the emissions that come from the polymer production stage. The emissions can be divided into energy-related emission and processing emission. Processing emission at the site is small compared to energy-related ones. It should be noted that plastics do not break down in landfills. Two solutions have been used over the last few years. The first is to burn the plastic that leads to toxic material into the air. This method is most commonly used today because plastics are petroleum based and have high heating properties. Countries like Japan and England have laws limiting the amount of petroleum-based products that can be incinerated and are moving toward more methods that recycle the product. Due to this trend, more effort is placed in the design phase of projects to ensure the correct mixture of recycled plastics, new polymer material, and additives for product performance. The second method is to recycle the plastic and make it into other products. This second solution can be used only for one of the two types of plastics. Thermoplastics can be melted, while thermoset plastics cannot be melted and have to be scrapped if a product is defective or at its end-of-life cycle. One area that has large opportunity for recycling is the plastic in automobiles. Current U.S. methods of recycling cars focus only on reusing the metal components. The plastic products are considered scrap and sent to landfills.

If we compare injection molding to other conventional manufacturing processes, injection molding appears to be on the same order of magnitude in term of energy consumption. For example, processes such as sand and die casting have similar energy requirements (11–15 MJ/kg). However, when compared to processes used in the semiconductor industry, the impact of injection molding seems significant. But in order to understand this point, we have to understand the product's widespread effect on

the economy. Injection-molding processes are more widely used and are growing in countries like China and India.

Although waste material is low and low levels of coolant are used in the process, the amount of energy used in the process has resulted in the research and development of ways to make the process more benign. It is critical to continue to improve the efficiency of the process in order to reduce the impact on the environment. It is essential to make a process that uses less energy, especially at this time when energy prices continue to rise.

5 MANUFACTURED PRODUCT

Most of the discussion in this chapter has focused on ways to ensure that manufacturing processes are environmentally benign. Any company that is morally and ethically committed to the goals of environmentally benign manufacturing cannot scrutinize its manufacturing processes without first giving due consideration to the manufactured product itself. It could legitimately be argued that the energy expenditure of certain products will easily surpass any savings in environmental impact achieved through optimally designed manufacturing processes very early in the product life cycle. An example is the large gas-guzzling truck or automobile, which is manufactured with the quaint notion of “bowing to customer demand” for large vehicles despite their poor fuel mileage performance.

It is curious, though, that a considerable marketing budget is expended to cultivate this customer demand. It is also curious that an automobile manufacturer recently withdrew from the marketplace a plug-in, all-electric vehicle that had managed to gain a great deal of approval from its customers. Yet, manufacturers offer the excuse that they cannot act unilaterally without suffering competitively in the marketplace. Lawmakers, too, are prone to succumb to the notion that “people should be free to buy the products they want.” Where does that leave the premise, or promise, of environmentally benign manufacturing? And where does that leave future generations, who are predestined to live in the environment we leave them?

REFERENCES

1. M. E. Zohdi, W. E. Biles, and D. B. Webster, “Production Processes and Equipment,” in M. Kutz (Ed.), *Mechanical Engineers’ Handbook: Book 3—Manufacturing and Management*, Wiley, New York, 2006, pp. 173–244.
2. M. E. Zohdi and W. E. Biles, “Metal Forming, Shaping and Casting,” in M. Kutz (Ed.), *Mechanical Engineers’ Handbook: Book 3—Manufacturing and Management*, Wiley, New York, 2006, pp. 245–285.
3. S. Javed, C. W. Lovell, and L. E. Wood. “Waste Foundry Sand in Asphalt Concrete,” in *Transportation Research Record 1437*, Transportation Research Board, Washington, DC, 1994.
4. American Foundry Society (AFS), “The AFS Teams with DOE, DOD, EPA, and DOT to Deliver Results for America,” AFS, Schaumburg, IL, 1999.
5. E. P. DeGarmo, J. T. Black, R. A. Kohser, and B. E. Klamecki, *Materials and Processes in Manufacturing*, 9th ed., Wiley, New York, 2003, pp. 920–998.
6. J.-M. Charrier, *Polymeric Materials and Processing: Plastics, Elastomers and Composites*, Hanser, Munich, 1990.

BIBLIOGRAPHY

- C. J. Backhouse, A. J. Clegg, and T. Staikos, “Reducing the Environmental Impacts of Metal Castings through Life-Cycle Management,” Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University, Loughborough, UK, *Progress in Industrial Ecology*, **1** (1–3), 2004.

- W. E. Biles, "Plastics Parts Processing Part I," in M. Kutz (Ed.), *Handbook of Materials Selection*, Wiley, New York, 2002, pp. 969–992.
- S. Dalquist and T. Gutowski, "Life-Cycle Analysis of Conventional Manufacturing Techniques," in *Massachusetts Institute of Technology Proceedings of IMECE2004: 2004 ASME International Mechanical Engineering Congress & Exposition*, Anaheim, CA, November 13–19, 2004.
- T. Gutowski, "Casting," available: www.geocities.com/fimutp/casting.pdf.
- D. O. Harper, "Plastics Parts Processing Part II," in M. Kutz (Ed.), *Handbook of Materials Selection*, Wiley, New York, 2002, pp. 993–1036.
- D. W. Richerson, "The Metal Casting Industry," Chapter 8, available: www.ms.ornl.gov/programs/energyeff/cfcc/iof/chap8.pdf.

CHAPTER 3

PRODUCTION PLANNING

Bhaba R. Sarker, Dennis B. Webster, and Thomas G. Ray
Louisiana State University
Baton Rouge, Louisiana

1 INTRODUCTION	53	5.3 Beyond Materials Requirements Planning	86
2 FORECASTING	54	6 JOB SEQUENCING AND SCHEDULING	87
2.1 General Concepts	54	6.1 Structure of the General Sequencing Problem	87
2.2 Qualitative Forecasting	55	6.2 Single-Machine Problem	88
2.3 Causal Methods	58	6.3 Flow Shops	90
2.4 Methods of Analysis of Time Series	60	6.4 Job Shops	92
2.5 Forecasting Error Analysis	61	6.5 Heuristics/Priority Dispatching Rules	94
2.6 Conclusions on Forecasting	63	6.6 Assembly Line Balancing	97
3 INVENTORY MODELS	63	7 JAPANESE MANUFACTURING PHILOSOPHY	104
3.1 General Discussion	63	7.1 Just-in-Time Philosophy/Kanban Mechanism	104
3.2 Types of Inventory Models	65	7.2 Time-based Competition	107
3.3 The Modeling Approach	66	8 SUPPLY CHAIN MANAGEMENT	108
4 AGGREGATE PLANNING—MASTER SCHEDULING	73	8.1 Distribution Logistics	108
4.1 Alternative Strategies to Meet Demand Fluctuations	74	8.2 Applications of Kanban Mechanism to Supply Chain	109
4.2 Aggregate Planning Costs	74	8.3 General Remarks	110
4.3 Approaches to Aggregate Planning	74	REFERENCES	111
4.4 Levels of Aggregation and Disaggregation	76	BIBLIOGRAPHY	113
4.5 Aggregate Planning Dilemma	76		
5 MATERIALS REQUIREMENTS PLANNING	77		
5.1 Procedures and Required Inputs	78		
5.2 Lot Sizing Techniques	85		

1 INTRODUCTION

The more manufacturing changes, the more it stays the same. Certainly the rapid introduction of newer technology and better approaches to management have led to unprecedented increases in productivity, but at its heart, the objective of manufacturing is still to provide the right product in the right quantity at the right time with the right quality at the right price to its customers.

The topics discussed in this chapter are related to how manufacturing organizations strive to meet this objective. Forecasting provides the manufacturer with a basis for anticipating

customer demand so as to have adequate product on hand when it is demanded. Of course, the preferred approach would be to wait for an order and then produce and ship immediately when the order arrives. This approach is, for practical purposes, impossible for products with any significant lead time in manufacturing, raw materials, or component supply. Consequently, most manufacturing facilities develop raw materials, in-process, and finished goods inventories, which have to be established and managed.

Aggregate planning approaches establish overall production requirements, and materials requirements planning techniques provide a methodology for ensuring that adequate inventory is available to complete the work required on products needed to meet forecasted customer demands. Job-sequencing methodologies are used to develop shop schedules for production processes to reduce the time for manufacturing products or meet other performance objectives. Although the basic functions in manufacturing are the same, the manner in which they are implemented drastically affects the effectiveness of the outcome. Improvements in technology have greatly increased the productivity that is achieved. Likewise, significant increases in reduction of costs and improved customer service are being achieved by changing management philosophies. Flowing from the *just-in-time* (JIT) concepts developed by the Japanese are practices designated as lean and/or agile manufacturing and the use of enterprise resource planning. More recently, the term supply chain management has been used to describe the inherent linkages among all of the functions of a manufacturing enterprise. The following materials are presented to introduce the reader to what these terms mean and how these approaches are related.

2 FORECASTING

2.1 General Concepts

The function of production planning and control is based on establishing a plan, revising the plan as required, and adhering to the plan to accomplish desired objectives. Plans are based on a forecast of future demand for the related products or services. Good forecasts are a requirement for a plan to be valid and functionally useful. When managers are faced with forecasts, they need to plan what actions must be taken to meet the requirements of the forecast. The actions taken thus prepare the organization to cope with the anticipated future state of nature that is predicated upon the forecast.¹⁻³

Forecasting methods are traditionally grouped into one of three categories: qualitative techniques, time-series analysis, or causal methods. The qualitative techniques are normally based on opinions or surveys. The basis for time-series analysis is historical data and the study of trends, cycles, and seasons. Causal methods are those that try to find relationships between independent and dependent variables, determining which variables are predictive of the dependent variable of concern. The method selected for forecasting must relate to the type of information available for analysis.

Definitions

Deseasonalization: The removal of seasonal effects from the data for the purpose of further study of the residual data.

Error analysis: The evaluation of errors in the historical forecasts done as a part of forecasting model evaluation.

Exponential smoothing: An iterative procedure for the fitting of polynomials to data for use in forecasting.

Forecast: Estimation of a future outcome.

Horizon: A future time period or periods for which a forecast is required.

Index number: A statistical measure used to compare an outcome, which is measured by a cardinal number with the same outcome in another period of time, geographic area, profession, etc.

Moving average: A forecasting method in which the forecast is an average of the data for the most recent n periods.

Qualitative forecast: A forecast made without using a quantitative model.

Quantitative forecast: A forecast prepared by the use of a mathematical model.

Regression analysis: A method of fitting a mathematical model to data by minimizing the sums of the squares of the data from a theoretical line.

Seasonal data: Data that cycle over a known seasonal period such as a year.

Smoothing: A process for eliminating unwanted fluctuations in data, which is normally accomplished by calculating a moving average or a weighted moving average.

Time-series analysis: A procedure for determining a mathematical model for data that are correlated with time.

Time-series forecast: A forecast prepared with a mathematical model from data that are correlated with time.

Trend: Underlying patterns of movement of historic data that becomes the basis for prediction of future forecasts.

2.2 Qualitative Forecasting

These forecasts are normally used for purposes other than production planning. Their validity is more in the area of policy making or in dealing with generalities to be made from qualitative data. Among these techniques are the Delphi method, market research, consensus methods, and other techniques based on opinion or historical relationships other than quantitative data. The Delphi method is one of a number of nominal group techniques. It involves prediction with feedback to the group that gives the predictor's reasoning. Upon each prediction the group is again polled to see if a consensus has been reached. If no common ground for agreement has occurred, the process continues moving from member to member until agreement is reached. Surveys may be conducted of relevant groups and their results analyzed to develop the basis for a forecast. One group appropriate for analysis is customers. If a company has relatively few customers, this select number can be an effective basis for forecasting. Customers are surveyed and their responses combined to form a forecast. Many other techniques are available for nonquantitative forecasting. An appropriate area to search if these methods seem relevant to a subjective problem at hand is the area of *nominal group techniques*. Quantitative forecasting involves working with numerical data to prepare a forecast. This area is further divided into two subgroups of techniques according to the data type involved. If historical data are available and it is believed that the dependent variable to be forecast relates only to time, an approach called time-series analysis is used. If the data available suggest relationships of the dependent variable to be forecast to one or more independent variables, then the techniques used fall into the category of causal analysis. The most commonly used method in this group is regression analysis.

Moving Average

A moving average can normally be used to remove the seasonal or cyclical components of variation. This removal is dependent on the choice of a moving average that contains sufficient data points to bridge the season or cycle. For example, a five-period centered moving average should be sufficient to remove seasonal variation from monthly data. A disadvantage to the use

of moving averages is the loss of data points due to the inclusion of multiple points into the calculation of a single point.

Example 1 Computation of Moving Average. A 5-year *simple* moving average in column 4 of Table 1 represents the forecast based on data of five recent past periods. Note that the 5-year *centered* moving average lost four data points—two on each end of the data series (last two columns in Table 1). Observation of the moving average indicated a steady downward trend in the data. The raw data had fluctuations that might tend to confuse an observer initially due to the apparent positive changes from time to time.

Weighted Moving Average

A major disadvantage of the moving average method—the effect of extreme data points— can be overcome by using a weighted moving average for N periods (Table 2). In this average the effect of the extreme data points may be decreased by weighing them less than the data points at the center of the group. There are many ways for this to be done.

Table 1 Moving-Average Computation

Year	Data	5-Year Moving Total	5-Year Moving Average	5-Year Centered Moving Total	5-Year Centered Moving Average
1	60.0	—	—	—	—
2	56.5	—	—	—	—
3	53.0	—	—	275.3	55.06
4	54.6	—	—	269.2	53.24
5	51.2	—	—	261.1	52.20
6	53.9	275.3	55.06	257.2	51.44
7	48.4	269.2	53.84	250.9	50.18
8	49.1	261.1	52.20	242.1	48.42
9	48.3	257.2	51.44	232.8	46.56
10	42.4	250.9	50.18	—	—
11	44.6	242.1	48.42	—	—

Table 2 Weighted Moving Average

Year	Data	5-Year Moving Total	5-Year Total Less the Center Value	Weighted Average $0.5(\text{Col 4})/4 + 0.5(\text{Col 2})$
1	60.0	—	—	—
2	56.5	—	—	—
3	53.0	275.3	222.3	54.3
4	54.6	269.2	214.6	54.1
5	51.2	261.1	209.9	51.8
6	53.9	257.2	203.3	52.4
7	48.4	250.9	202.5	49.5
8	49.1	242.1	193.0	48.7
9	48.3	232.8	184.5	47.2
10	42.4	—	—	—
11	44.6	—	—	—

Table 3 Forecasts by Simple Moving-Average and Weighted Moving-Average Methods

Period	Moving-Average Forecast	Weighted Average Forecast
3	55.1	54.3
4	53.8	54.1
5	52.2	51.8
6	51.4	52.4
7	50.2	49.5
8	48.4	48.7
9	46.6	47.2

One method would be to weigh the center point of the N -period (in this case a five-period) average as 50% of the total with the remaining points weighted for the remaining 50%. For $N = 5$, the total is $60.0 + 56.6 + 53.0 + 54.6 + 51.2 = 275.3$, and the five-period total less the centered value is $275.3 - 53.0 = 222.2$. Hence, weighted average is $0.5(222.3/4) + 0.5(53.0) = 54.3$. Similar to the calculation for Example 1, this would yield the results as shown in Table 2.

Example 2 Weighted Moving Average. Table 3 displays the two forecasts. The results are comparable to the weighted-average forecast, distinguishing a slight upswing from period 5 to period 6, which was ignored by the moving-average method.

Exponential Smoothing

This method determines the forecast (F) for the next period as the weighted average of the last forecast and the current demand (D). The current demand is weighted by a constant, α , and the last forecast is weighted by the quantity $1 - \alpha$ ($0 \leq \alpha \leq 1$). New forecast = α (demand for current period) + $(1 - \alpha)$ forecast for current period. This can be expressed symbolically as

$$F_t = \alpha D_{t-1} + (1 - \alpha)F_{t-1} \quad (1)$$

The forecast F_t is the one-step ahead forecast for the period t made in period $t - 1$. Using the similarity as in Eq. (1), we can write

$$F_{t-1} = \alpha D_{t-2} + (1 - \alpha)F_{t-2} \quad (2)$$

Substituting Eq. (2) in Eq. (1), we have

$$F_t = \alpha D_{t-1} + \alpha(1 - \alpha)D_{t-2} + (1 - \alpha)^2 F_{t-2} \quad (3)$$

$$F_t = \alpha D_{t-1} + \alpha(1 - \alpha)D_{t-2} + \alpha(1 - \alpha)^2 D_{t-3} + (1 - \alpha)^3 F_{t-3} \quad (4)$$

and, in general,

$$F_t = \sum_{i=0}^{\infty} \alpha(1 - \alpha)^i D_{t-i-1} = \sum_{i=0}^{\infty} \alpha^i D_{t-i-1} \quad (5)$$

Obviously, the exponential weights are $\alpha(1 - \alpha)^i = a_0 > a_1 > a_2 > \dots > a_{i-1} > a_i$, where $\sum_{i=1}^{\infty} a_i = \sum_{i=1}^{\infty} \alpha(1 - \alpha)^i = 1$, indicating that the exponential smoothing technique applies a declining set of weights to all past data. For further treatment of the exponential smoothing techniques, readers may refer to Nahmias⁴ or Bedworth and Bailey.¹

Normally, the forecast for the first period is taken to be the actual demand for that period (i.e., forecast and demand are the same for the initial data point). The smoothing constant is

Table 4 Forecasts F_t for Various α Values by Exponential Method

Period	Demand	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
1	85	85.0	85.0	85.0
2	102	85.0	85.0	85.0
3	110	86.7	88.4	90.1
4	90	89.0	92.7	96.1
5	105	89.1	92.2	94.3
6	95	90.7	94.8	97.5
7	115	91.1	94.8	96.8
8	120	93.5	98.8	102.3
9	80	96.2	103.0	107.6
10	95	94.6	98.4	99.3

chosen as a result of analysis of error by a method such as mean absolute deviation coupled with the judgment of the analyst. A high value of α makes the forecast very responsive to the occurrence in the last period. Similarly, a small value would lead to a lack of significant response to the current demand. Evaluations must be made in light of the cost effects of the errors to determine what value of α is best for a given situation. Example 3 shows the relationship between actual data and forecasts for various values of α .

Example 3 Exponential Smoothing. Table 4 provides the forecasts by exponential smoothing method using different values of α .

2.3 Causal Methods

This category of methods falls within the second group of quantitative forecasting methods mentioned earlier. These methods assume that there are certain factors that have a cause–effect relationship with the outcome of the quantity to be forecast and that knowledge of these factors will allow a more accurate prediction of the dependent quantity. The statistical models of regression analysis fall within this category of forecasting.

Basic Regression Analysis

The simplest model for regression analysis is the linear model. The basic approach involves the determination of a theoretical line that passes through a group of data points that appear to follow a linear relationship. The desire of the modeler is to determine the equation for the line that would minimize the sums of the squares of the deviations of the actual points from the corresponding theoretical points. The values for the theoretical points are obtained by substituting the values of the independent variable, Y_i , into the functional relationship

$$\hat{Y}_i = a + bx_i \quad (6)$$

The difference between the data and the forecasted value of point i is

$$e_i = Y_i - \hat{Y}_i \quad (7)$$

Squaring this value and summing the relationship over the N related points yields the total error, E , as

$$E = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \quad (8)$$

Substituting the functional relationship for the forecasted value of Y gives

$$E = \sum_{i=1}^N (Y_i - \hat{a} - \hat{b}x_i)^2 \quad (9)$$

By using this relationship, taking the partial derivatives of E with respect to a and b , and solving the resulting equations simultaneously, we obtain the normal equations for least squares for the linear regression case:

$$\begin{aligned} \sum Y &= aN + b \sum X \\ \sum XY &= a \sum X + b \sum X^2 \end{aligned}$$

Solving these equations yields values for a and b :

$$b = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2} \quad (10)$$

and

$$a = \bar{Y} - b\bar{X} \quad (11)$$

The regression equation is then $Y_i = a + bx_i$, and the correlation coefficient, r , which gives the relative importance of the relationship between x and y , is

$$r = \frac{N \sum XY - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}} \quad (12)$$

This value of r can range from +1 to -1. The plus sign would indicate a positive correlation (i.e., large values of x are associated with large values of y ; a negative correlation implies that large values of x are associated with small values of y) and the negative sign would imply a negative correlation.

Example 4 Simple Linear Regression. From the X and Y data set in Table 5, and using the Eqs. (10)–(12), the following computational results yield

$$b = \frac{3(47) - 9(15)}{3(29) - 81} = 1 \quad \text{and} \quad a = 5 - 1(3) = 2$$

and the linear forecast model is $Y = 2 + X$. The correlation coefficient is

$$r = \frac{3(47) - 9(15)}{\sqrt{3(29) - 81} \sqrt{3(77) - 225}} = 1 \quad (13)$$

which indicates that the X - Y data are 100% positively correlated.

Table 5 Linear Forecasting

	Y	X	XY	X^2	Y^2
	4	2	8	4	16
	5	3	15	9	25
	6	4	24	16	36
Σ	15	9	47	29	77
Mean	5	3	—	—	—

Quadratic Regression

This regression model is used when the data appear to follow a simple curvilinear trend and the fit of a linear model is not adequate. The procedure for deriving the normal equations for quadratic regression is similar to that for linear regression. The quadratic model has three parameters that must be estimated, however: the constant term, a , the coefficient of the linear term, b , and the coefficient of the square term, c . The model is

$$Y_i = a + bx_i + cx_i^2 \quad (14)$$

Its normal equations are

$$\begin{aligned} \sum Y &= Na + b \sum X + c \sum X^2 \\ \sum XY &= a \sum X + b \sum X^2 + c \sum X^3 \\ \sum X^2 Y &= a \sum X^2 + b \sum X^3 + c \sum X^4 \end{aligned}$$

The normal equations for least squares for a cubic curve, quartic curve, etc. can be generalized from the expressions for the linear and quadratic models.

2.4 Methods of Analysis of Time Series

We now discuss in general several methods for analysis of time series. These methods provide ways of removing the various components of the series, isolating them, and providing information for their consideration should it be desired to reconstruct the time series from its components.

The movements of a time series are classified into four types: long-term or *trend* movements, *cyclical* movements, *seasonal* movements, and *irregular* movements. Each of these components can be isolated or analyzed separately. Various methods exist for the analysis of the time series. These methods decompose the time series into its components by assuming that the components are either multiplicative or additive. If the components are assumed to be multiplicative, the following relationship holds:

$$Y = T \times C \times S \times I \quad (15)$$

where Y is the outcome of the time series, T is the trend value of the time series, and C , S , and I are indices, respectively, for cyclical, seasonal, and irregular variations.

To process data for this type of analysis it is best to first plot the raw data to observe its form. If the data are yearly, they need no deseasonalization. If they are monthly or quarterly data, they can be converted into yearly data by summing the data points that would add to a year before plotting. (Seasonal index numbers can be calculated to seasonalize the data later if required.) By plotting yearly data, the period of apparent data cycles can be determined or approximated. A centered moving average of appropriate order can be used to remove the cyclical effect in the data. Further, cyclical indices can be calculated when the order of the cycle has been determined. At this point the data contain only the trend and irregular components of variation. Regression analysis can be used to estimate the trend component of the data, leaving only the irregular, which is essentially a forecasting error.

Index numbers are calculated by grouping data of the same season together, calculating the average over the season for which the index is to be prepared, and then calculating the overall average of the data over each of the seasons. Once the seasonal and overall averages are obtained, the seasonal index is determined by dividing the seasonal average by the overall average.

Table 6 Sales Data for Two Years

Month	Year 1	Year 2	Total	Average	Index
January	20	24	44	22.0	0.904
February	23	27	50	25.0	1.026
March	28	30	58	29.0	1.191
April	32	35	67	33.5	1.375
May	35	36	71	35.5	1.456
June	26	28	54	27.0	1.117
July	25	27	52	26.0	1.066
August	23	23	46	23.0	0.944
September	19	17	36	18.0	0.737
October	21	22	43	21.5	0.882
November	18	19	37	18.5	0.750
December	12	14	26	13.5	0.552
TOTAL	282	302	584	292 ^a	12.014

^aMonthly average = 24.333.

Example 5 Average Forecast. A business has been operational for 24 months. The sales data in thousands of dollars for each of the monthly periods are given in Table 6. The overall total is $282 + 302 = 584$ and the average is $584/24 = 24.333$. The index for January would be

$$I_{\text{Jan}} = \frac{0.5(20 + 24)}{24.333} = 0.904$$

For the month of March the index would be

$$I_{\text{Mar}} = \frac{0.5(28 + 30)}{24.333} = 1.191$$

To use the index, a trend value for the year's sales would be calculated, the average monthly sales would be obtained, and then this figure would be multiplied by the index for the appropriate month to give the month's forecast.

Note that a season can be defined as any period for which data are available for appropriate analysis. If there are seasons within a month, i.e., 4 weeks in which the sales vary considerably according to a pattern, a forecast could be indexed within the monthly pattern also. This would be a second indexing within the overall forecast. Further, seasons could be chosen as quarters rather than months or weeks. This choice of the period for the analysis is dependent on the requirements for the forecast.

Data given on a seasonal basis can be deseasonalized by dividing them by the appropriate seasonal index. Once this has been done they are labeled *deseasonalized data*. These data still contain the trend, cyclical, and irregular components after this adjustment.

Seasonal Forecasts

Given a set of quarterly data for 3 years as given in Table 7, four more tables (Tables 7a–7d) are generated for seasonal and yearly average, seasonal indices, deseasonalized forecast, and seasonalized forecasts.

2.5 Forecasting Error Analysis

One common method of evaluating forecast accuracy is termed mean absolute deviation (MAD) from the procedure used in its calculation. For each available data point a comparison of the forecasted value is made to the actual value. The absolute value of the differences is calculated.

Table 7 Seasonal Sales Data

Year	Q_1	Q_2	Q_3	Q_4
1	520	730	820	530
2	590	810	900	600
3	650	900	1000	650

Table 7a Seasonal Data

Year	Q_1	Q_2	Q_3	Q_4	Total	Average
1	520	730	820	530	2600	650
2	590	810	900	600	2900	725
3	650	900	1000	650	3200	800
Total	1760	2440	2720	1780	—	—
Average	586.0	813.6	906.7	593.3	—	—

Table 7b Seasonal Index

Year	Q_1	Q_2	Q_3	Q_4	Total	Average
1	0.800 ^a	1.123	1.261	0.815	4.0	
2	0.813 ^b	1.117	1.241	0.828	4.0	
3	0.812	1.125	1.250	0.812	4.0	
Total	2.425	3.365	3.752	2.455	—	
Average	0.808	1.122	1.251	0.818	4.0	

^a520/650 = 0.800. ^b590/725 = 0.813.

Table 7c Deseasonalized Data (Unadjusted) and Forecast

Year, t	Q_1	Q_2	Q_3	Q_4	Total	Average
1	643 ^a	650 ^c	655	648	2596	649
2	730 ^b	720	719	733	2902	725
3	804	802	799	794	3199	800
4						876.67
5	$F_t = 572.67 + 76t$					952.57

^a520/0.808 = 643. ^b590/0.808 = 730. ^c730/1.125 = 650.

Table 7d Reseasonalized Forecast (Adjusted) from Deseasonalized Data

Year, t (Index)	Q_1 (0.808)	Q_2 (1.122)	Q_3 (1.251)	Q_4 (0.818)	Total	Average
4	707 ^a	982 ^c	1095	716	—	876.67
5	769 ^b	1068	1191	778	—	952.57

^a876.67 × 0.808 = 707. ^b952.67 × 0.808 = 769. ^c876.67 × 1.122 = 982.

This absolute difference is then summed over all values and its average calculated to give the evaluation:

$$\text{MAD} = \frac{\text{Sum of the absolute deviations}}{\text{Number of deviations}} = \frac{1}{N} \sum |Y_i - \hat{Y}_i| \quad (16)$$

Alternative forecasts can be analyzed to determine the value of MAD and a comparison made using this quantity as an evaluation criteria. Other criteria can also be calculated. Among these are the mean square of error (MSE) and the standard error of the forecast (S_{xy}). These evaluation criteria are calculated as

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \quad (17)$$

and

$$S_{xy} = \frac{\sqrt{\sum Y^2 - a \sum Y - a \sum XY}}{N - 2} \quad (18)$$

In general, these techniques are used to evaluate the forecast, and then the results of the various evaluations together with the data and forecasts are studied. Conclusions may then be drawn as to which method is preferred or the results of the various methods compared to determine what they in effect distinguish.

2.6 Conclusions on Forecasting

A number of factors should be considered in choosing a method of forecasting. One of the most important factors is *cost*. The problem of valuing an accurate forecast is presented. If the question “how will the forecast help and in what manner it will save money?” can be answered, a decision can be made regarding the allocation of a percentage of the savings to the cost of the forecasting process. Further, concern must be directed to the required *accuracy* of a forecast to achieve desired cost reductions. Analysis of past data and the testing of the proposed model using this historical data provide a possible scenario for hypothetical testing of the effects of cost of variations of actual occurrences from the plan value (forecast).

In many cases an inadequate database will prohibit significant analysis. In others the database may not be sufficient for the desired projection into the future. The answers to each of these questions are affected by the type of product or service for which the forecast is to be made as well as the value of the forecast to the planning process.

3 INVENTORY MODELS

3.1 General Discussion

Normally, items waiting to be purchased or sold are considered to be in inventory. One of the most pressing problems in the manufacturing and sale of goods is the control of this inventory. Many companies experience financial difficulties each year due to a lack of adequate control in this area. Whether it is raw material used to manufacture a product or products waiting to be sold, problems arise when too many or too few items are available. The greatest number of problems arises when too many items are held in inventory.

The primary factor in the reduction of inventory costs is deciding when to order, how much to order, and if back-ordering is permissible. Inventory control involves decisions by management as to the source from which the inventory is to be procured and as to the quantity to be procured at the time. This source could be from another division of the company handled as an intrafirm transfer, outside purchase from any of a number of possible vendors, or manufacture of the product in house.

The basic decisions to be made once a source has been determined are how much to order and when to order. Inherent in this analysis is the concept of demand. Demand can be known or unknown, probabilistic or deterministic, constant or lumpy. Each of these characteristics affects the method of approaching the inventory problem.

For the *unknown demand* case a decision must be made as to how much the firm is willing to risk. Normally, the decision would be to produce some k units for sale and then determine after some period of time to produce more or to discontinue production due to insufficient demand. This amounts to the reduction of the unknown demand situation to one of a lumpy demand case after the decision has been made to produce the batch of a finite size. Similarly, if a decision is made to begin production at a rate of n per day until further notice, the unknown demand situation has been changed to a constant known demand case.

Lumpy demand, or demand that occurs periodically with quantities varying, is frequently encountered in manufacturing and distribution operations. It is distinguished from the *known demand* case. This second case is that of a product that has historic data from which forecasts of demand can be prepared. A factor of concern in these situations is the lead time and the unit requirement on a periodic basis. The following are the major factors to be considered in the modeling of the inventory situation.

Demand is the primary stimulus on the procurement and inventory system and it is the justification for its existence. Specifically, the system may exist to meet the demand of customers, the spare parts demand of an operational weapons system, the demand of the next step in a manufacturing process, etc. The characteristic of demand, although independent of the source chosen to replenish inventories, will depend on the nature of the environment giving rise to the demand.

The simplest demand pattern may be classified as deterministic. In this special case, the future demand for an item may be predicted with certainty. Demand considered in this restricted sense is only an approximation of reality. In the general case, demand may be described as a random variable that takes on values in accordance with a specific probability distribution.

Procurement quantity is the order quantity, which in effect determines the frequency of ordering and is related directly to the maximum inventory level.

Maximum shortage is also related to the inventory level.

Item cost is the basic purchase cost of a unit delivered to the location of use. In some cases delivery cost will not be included if that cost is insignificant in relation to the unit cost. (In these cases the delivery cost will be added to overhead and not treated as a part of direct material costs.)

Holding costs are incurred as a function of the quantity on hand and the time duration involved. Included in these costs are the real out-of-pocket costs, such as insurance, taxes, obsolescence, and warehouse rental and other space charges, and operating costs, such as light, heat, maintenance, and security. In addition, capital investment in inventories is unavailable for investment elsewhere. The rate of return forgone represents a cost of carrying inventory.

The inventory holding cost per unit of time may be thought of as the sum of several cost components. Some of these may depend on the maximum inventory level incurred. Others may depend on the average inventory level. Still others, like the cost of capital invested, will depend on the value of the inventory during the time period. The determination of holding cost per unit for a specified time period depends on a detailed analysis of each cost component.

Ordering cost is the cost incurred when an order is placed. It is composed of the cost of time, materials, and any expense of communication in placing an order. In the case of a manufacturing model it is replaced by *setup cost*. Setup cost is the cost incurred when a machine's tooling or jigs and fixtures must be changed to accommodate the production of a different part or product.

Shortage cost is the penalty incurred for being unable to meet a demand when it occurs. This cost does not depend on the source chosen to replenish the stock but is a function of the number of units short and the time duration involved.

The specific dollar penalty incurred when a shortage exists depends on the nature of the demand. For instance, if the demand is that of customers of a retail establishment, the shortage cost will include the loss of goodwill. In this case the shortage cost will be small relative to the cost of the item. If, however, the demand is that of the next step of a manufacturing process, the cost of the shortage may be high relative to the cost of the item. Being unable to meet the requirements for a raw material or a component part may result in lost production or even closing of the plant. Therefore, in establishing shortage cost, the seriousness of the shortage condition and its time duration must be considered.⁵

3.2 Types of Inventory Models

Deterministic models assume that quantities used in the determination of relationships for the model are all known. These quantities are such things as demand per unit of time, lead time for product arrival, and costs associated with such occurrences as a product shortage, the cost of holding the product in inventory, and that cost associated with placing an order for product.

Constant demand is one case that can be analyzed within the category of deterministic models. It represents very effectively the case for some components or parts in an inventory that are used in multiple parents, these multiple parent components having a composite demand that is fairly constant over time.

Lumpy demand is varying demand that occurs at irregular points in time. This type of demand is normally a dependent demand that is driven by an irregular production schedule affected by customer requirements. Although the same assumptions are made regarding the knowledge of related quantities, as in the constant demand case, this type of situation is analyzed separately under the topic of materials requirements planning (MRP). This separation of methodology is due to the different inputs to the modeling process in that the knowledge about demand is approached by different methods in the two cases.

Probabilistic models consider the same quantities as do the deterministic models but treat the quantities that are not cost related as random variables. Hence, demand and lead time have their associated probability distributions. The added complexity of the probabilistic values requires that these models be analyzed by radically different methods.

Definitions of Terms

The following terms are defined to clarify their usage in the material related to inventory that follows. Where appropriate, a literal symbol is assigned to represent the term.

Inventory (*I*): Stock held for the purpose of meeting a demand either internal or external to the organization.

Lead time (*L*): The time required to replenish an item of inventory by either purchasing from a vendor or manufacturing the item in-house.

Demand (*D*): The number of units of an inventory item required per unit of time.

Reorder point (*r*): The point at which an order must be placed for the procured quantity to arrive at the proper time or, for the manufacturing case, the finished product to begin flowing into inventory at the proper time.

Reorder quantity (*Q*): The quantity for which an order is placed when the reorder point is reached.

Demand during lead time (D_L): This quantity is the product of lead time and demand. It represents the number of units that will be required to fulfill demand during the time that it takes to receive an order that has been placed with a vendor.

Replenishment rate (P): This quantity is the rate at which replenishment occurs when an order has been placed. For a purchase situation it is infinite (when an order arrives, in an instant the stock level rises from 0 to Q). For the manufacturing situation it is finite.

Shortage: The units of unsatisfied demand that occur when there is an out-of-stock situation.

Back-order: One method of treating demand in a shortage situation when it is acceptable to the customer. (A notice is sent to the customer saying that the item is out of stock and will be shipped as soon as it becomes available.)

Lumpy demand: Demand that occurs in an aperiodic manner for quantities whose volume may or may not be known in advance. Constant demand models should normally never be used in a lumpy demand situation. The exception would be a component that is used for products that experience lumpy demand but itself experiences constant demand. The area of MRP was developed to deal with the lumpy demand situations.

3.3 The Modeling Approach

Modeling in operations research involves the representation of reality by the construction of a model in one of several alternative ways. These models may be iconic, symbolic, or mathematical. For inventory models the latter is normally the selection of choice. The model is developed to represent a concept whose relationships are to be studied. As much detail can be included in a particular model as is required to effectively represent the situation. The detail omitted must be of little significance to its effect on the model. The model's fidelity is the extent to which it accurately represents the situation for which it is constructed.

Inventory modeling involves building mathematical models to represent the interactions of the variables of the inventory situation to give results adequate for the application at hand. In this section treatment is limited to deterministic models for inventory control. Probabilistic or stochastic models may be required for some analysis. References 2, 5, and 6 may be consulted if more sophisticated models are required.

General

Using the terminology defined above, a basic logic model of the general case inventory situation will be developed. The objective of inventory management will normally be to determine an operating policy that will provide a means to reduce inventory costs. To reduce costs a determination must first be made as to what costs are present. The general model is as follows:

$$\text{Total cost} = \left(\begin{array}{c} \text{Cost of} \\ \text{items} \end{array} \right) + \left(\begin{array}{c} \text{Cost of} \\ \text{ordering} \end{array} \right) + \left(\begin{array}{c} \text{Cost of holding} \\ \text{items in stock} \end{array} \right) + \left(\begin{array}{c} \text{Cost of} \\ \text{shortage} \end{array} \right)$$

This cost is stated without a base period specified. Normally it will be stated as a per-period cost with the period being the same period as the demand rate (D) period.

Models of Inventory Situations

Purchase Model with Shortage Prohibited. This model is also known as a infinite replenishment rate model with infinite storage cost. This latter name results from the slope of the replenishment rate line (it is vertical) when the order arrives. The quantity on hand instantaneously changes

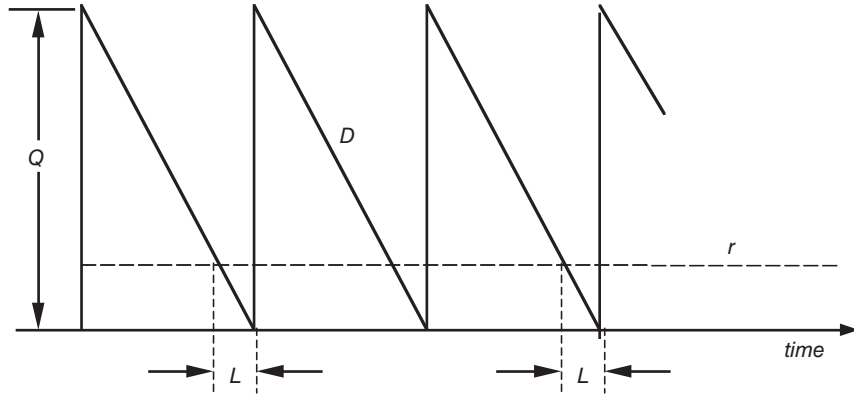


Figure 1 Basic inventory model with instantaneous replenishment.

from zero to Q . The shortage condition is preempted by the assignment of an infinite value to storage cost (see Fig. 1).

For this case, for unit purchase price of C_i dollars/unit, the item cost per period is symbolically $C_i D$. If the ordering cost is C_p dollars/order, the lot ordering cost is $(C_p D)/Q$. The shortage cost is zero since shortage is prohibited and the inventory holding cost is $(C_h Q)/2$. The equation for total cost (TC) is then

$$TC(Q) = C_i D + \left(\frac{D}{Q}\right) C_p + \left(\frac{Q}{2}\right) C_h \quad (19)$$

Analysis of this model reveals that the first component of cost, the cost of items, does not vary with Q . (Here we are assuming a constant unit cost; purchase discounts models are covered later.) The second component of cost, the cost of ordering, will vary on a per-period basis with the size of the order (Q). For larger values of Q , the cost will be smaller since fewer orders will be required to receive the fixed demand for the period. The third component of cost, cost of holding items in stock, will increase with increasing order size Q and conversely decrease with smaller order sizes. The fourth component of cost, cost of shortage, is affected by the reorder point. It is not affected by the order size and for this case shortage is not permitted.

This equation is essentially obtained by determining the cost of each of the component costs on a per cycle basis and then dividing that expression by the number of periods per cycle (Q/D). To obtain the extreme point(s) of the function, it is necessary to take the derivative of $TC(Q)$ with respect to Q , equate this quantity to zero, and solve for the corresponding value(s) of Q . This yields

$$0 = -\frac{C_p D}{Q^2} + \frac{C_h}{2}$$

or

$$\hat{Q} = \sqrt{\frac{2C_p D}{C_h}} \quad (20)$$

and

$$L = DT \quad (21)$$

Inspection of the sign of the second derivative of this function reveals that the extreme point is a minimum. This fits the objective of the model formulation. The quantity to be ordered at any point in time is then \hat{Q} and the time to place the order will be when the inventory level drops to r (the units consumed during the lead time for receiving the order).

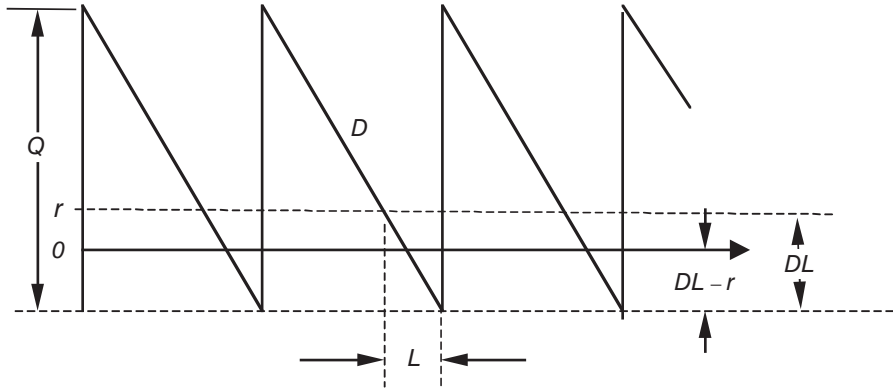


Figure 2 Model with shortage permitted.

Purchase Model with Shortage Permitted. This model is also known as an infinite replenishment rate model with finite shortage costs (see Fig. 2). For this model the product cost and the ordering cost are the same as for the previous model:

$$C_i D + \frac{C_p D}{Q}$$

The holding cost is different, however. It is given by

$$\frac{C_h [Q - (DL - r)]^2}{2Q}$$

This represents the unit periods of holding per cycle times the holding cost per unit period. The unit periods of holding is obtained from the area of the triangle whose altitude is $Q - (DL - r)$ and whose base is the same quantity divided by the slope of the hypotenuse. In the same manner the unit periods of shortage is calculated. For that case the altitude is $DL - r$ and the base is $DL - r$ divided by D . The shortage cost component is then

$$\frac{C_s (DL - r)^2}{2Q}$$

The total cost per period is given by

$$\begin{aligned} TC(Q, DL - r) = & C_i D + \left(\frac{D}{Q}\right) C_p + \left\{ \frac{[Q - (DL - r)]^2}{2Q} \right\} C_h \\ & + \left\{ \frac{[DL - r]^2}{2Q} \right\} C_s \end{aligned} \quad (22)$$

Note that the quantity $DL - r$ is used as a variable. This is done for the purposes of amplifying the equations that result when the partial derivatives are taken for the function. Taking these derivatives and solving the resulting equations simultaneously for the values of Q and $DL - r$ yields the following relationships:

$$\hat{Q} = \sqrt{\frac{2C_p D}{C_h} + \frac{C_p D}{C_s}} \quad (23)$$

$$\hat{L} = DL - \sqrt{\frac{2C_n C_p D}{C_s (C_h + C_s)}} \quad (24)$$

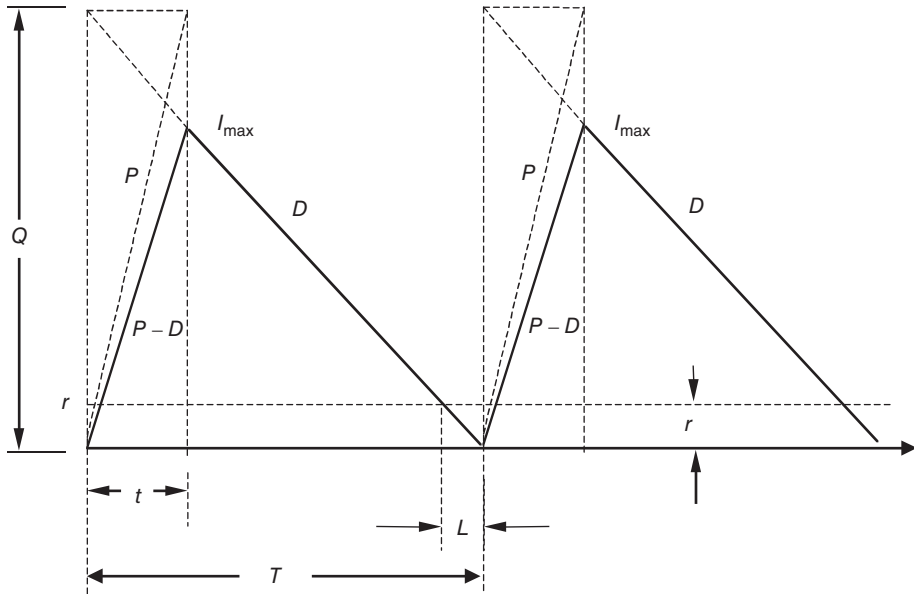


Figure 3 Manufacturing model with no shortage.

Manufacturing Model with Shortage Prohibited. This model is also known as a finite replenishment rate model with infinite storage costs. Figure 3 illustrates the situation, expressed as

$$\hat{Q} = \sqrt{\frac{2C_p D}{C_h(1 - D/P)}} \tag{25}$$

$$r = DL \tag{26}$$

Manufacturing Model with Shortage Permitted. This model is also known as a finite replenishment rate model with finite shortage costs. It is the most complex of the models treated here as it is the general case model. All of the other models can be obtained from it by properly defining the replenishment rate and shortage cost. For example, the purchase model with shortage prohibited is obtained by defining the manufacturing rate and the storage cost as infinite. Upon doing this, the equations reduce to those appropriate for the stated situation (see Fig. 4). For this model the expressions for \hat{Q} and \hat{r} are

$$\hat{Q} = \sqrt{\frac{1}{1 - D/P}} \sqrt{\frac{2C_p D}{C_h} + \frac{2C_p D}{C_s}} \tag{27}$$

$$\hat{r} = DL \sqrt{\frac{2C_p D(1 - D/P)}{C_s(1 + C_s/C_h)}} \tag{28}$$

Models for Purchase Discounts

MODELS FOR PURCHASE DISCOUNTS WITH FIXED HOLDING COST. In this situation, the holding cost (C_w) is assumed to be fixed, not a function of unit costs. A supplier offers a discount for ordering a larger quantity. The normal situation is as shown in Table 8.

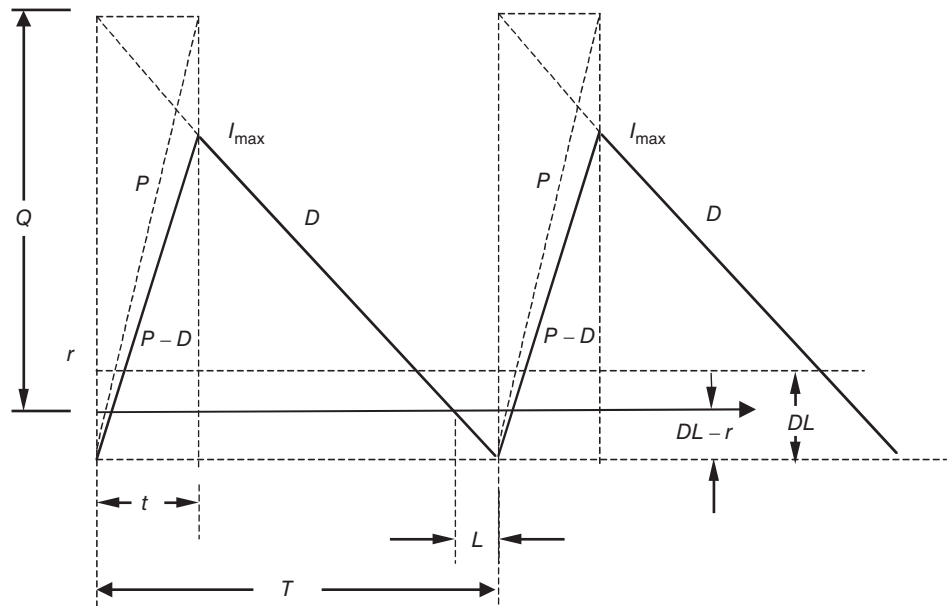


Figure 4 Manufacturing models with shortage permitted.

Table 8 Purchased Quantity for Different Price

Range of Quantity Purchased	Price, C_p
$1 - q_1$	C_1
$q_1 = 1 - q_2$	C_2
$q_2 = 1 - q_3$	C_3
—	—
$q_{m-1} = 1 - q_m$	C_m

The decision maker must apply the appropriate economic order quantity (EOQ) purchase model, either finite or zero shortage (infinite storage) cost. Upon choice of the appropriate model the following procedure will apply:

1. Evaluate Q and calculate $TC(\hat{Q})$.
2. Evaluate $TC(q_{k+1})$, where $q_{k=1}$ is the smallest quantity in the price break interval above that interval where q lies.
3. If $TC(\hat{q}) < TC(q_{k+1})$ the ordering quantity will be \hat{q} . If not, go to step 4.
4. Since the total cost of the minimum quantity in the next interval above that interval containing \hat{q} is a basic amount, an evaluation must be made successively of total costs of the minimum quantities in the succeeding procurement intervals until one reflects on increase in cost or the last choice is found to be the minimum.

Example 6 Discounted Inventory Model. In a situation where shortage is not permitted, the ordering cost is \$50, the holding cost is \$1 per unit year, and the demand is 10,000 units/year:

$$\hat{Q} = \sqrt{\frac{2C_p D}{C_w}} = \sqrt{\frac{2(\$50)10,000}{1}} = 1000 \text{ units}$$

$$TC(Q) = C_i D + \frac{C_h Q}{2} + \frac{D}{Q} C_p$$

$$TC(\hat{Q}) = \$20(10,000) + \$1 \left(\frac{1000}{2} \right) + \$50 \left(\frac{\$810,000}{1,000} \right) = \$201,000/\text{year}$$

The question is whether the smaller quantity in the next discount interval (1200–1799) gives a lower total cost (see Table 9):

$$TC(1200) = \$18(10,000) + \$1 \left(\frac{1200}{2} \right) + \$50 \left(\frac{10,000}{1800} \right) = \$181,033$$

Since this is a lower cost, an evaluation must be made of the smallest quantity in the next interval, 1800:

$$TC(1800) = 16.50(10,000) + \$1 \left(\frac{1800}{2} \right) + \$50 \left(\frac{10,000}{1800} \right) = \$166,175$$

Since there are no further intervals for analysis, this is the lowest total cost and its associated q , 1800, should be chosen as the optimal \hat{Q} . The total cost function for this model is shown in Fig. 5.

QUANTITY DISCOUNT MODEL WITH VARIABLE HOLDING COST. In this case, the holding cost is variable with unit cost, i.e., $C_w = KC_i$. Again, the appropriate model must be chosen for shortage conditions. For the zero shortage case

$$\hat{Q} = \sqrt{\frac{2C_p D}{KC_i}} \quad (29)$$

To obtain the optimal value of Q in this situation, the following procedure must be followed:

1. Evaluate \hat{Q} using the expression above and the item cost for the first interval.
 - a. If the value of \hat{Q} falls within the interval for C_i , use this $TC(\hat{Q})$ for the smallest cost in the interval.
 - b. If \hat{Q} is greater than the maximum quantity in the interval, use Q_{\max} , where Q_{\max} is the greatest quantity in the interval, and evaluate $TC(Q_{\max})$ as the lowest cost point in the interval.

Table 9 Discounted Price Range

Q	C_i
0–500	22.00
501–1199	20.00
1200–1799	18.00
1800–∞	16.50

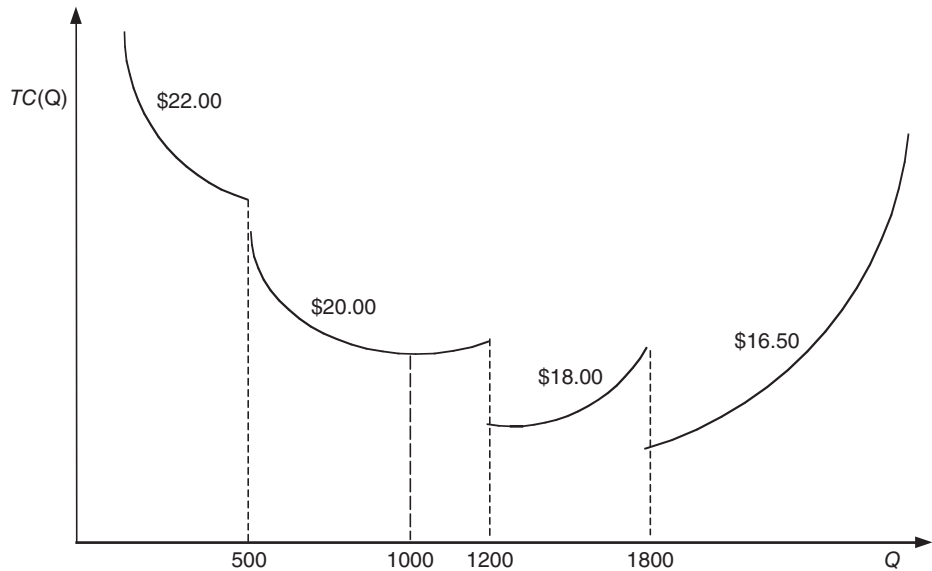


Figure 5 Total cost function for a quantity discount model with fixed holding cost.

- c. If \hat{Q} is less than the smallest Q in the interval, use Q_{\min} , where Q_{\min} is the smallest quantity in the interval, as the best quantity and evaluate $TC(Q_{\min})$.
2. For each cost interval follow the steps of part 1 of the procedure to evaluate best values in the interval.
3. Choose the minimum total cost from the applications of steps 1 and 2.

Example 7 Quantity Discounted Model with Variable Holding Cost. Using the same data as the Example 6, assume $C_h = 0.05C_1$:

$$\hat{Q} = \sqrt{\frac{2(50)10,000}{(0.05)(22)}} \cong 990 \text{ units/order}$$

Since $990 > 500$, the smallest cost in the interval will be at

$$\begin{aligned} TC(1000) &= 22.00(10,000) + \$1 \left(\frac{500}{2} \right) + \$50 \left(\frac{10,000}{500} \right) \\ &= 220,000 + 250 + 1000 = \$221,250 \end{aligned}$$

Using the second interval,

$$\hat{Q} = \sqrt{\frac{2(50)10,000}{0.05(20)}} = 1000 \text{ units/order}$$

Since this value falls within the interval, $TC(1000)$ is calculated

$$TC(1000) = 20(10,000) + \$1 \left(\frac{1000}{2} \right) + 50 \left(\frac{10,000}{1000} \right) = \$201,000$$

For the next interval (1201–1799), $C_3 = 18.00$ and

$$\hat{Q} = \sqrt{\frac{2(50)10000}{0.05(18.00)}} \approx 1050 \text{ units/order}$$

(which falls outside the interval to the left). Hence, the smallest quantity in the interval will be used. This is lower than either interval previously evaluated. For the next interval

$$\hat{Q} = \sqrt{\frac{2(50)10,000}{0.05(16.50)}} \cong 1101 \text{ units}$$

Hence, the smallest quantity in the interval must be used:

$$\begin{aligned} \text{TC}(1800) &= 16.50(10,000) + \$1 \left(\frac{1800}{2} \right) + \$50 \left(\frac{10,000}{500} \right) \\ &= 165,000 + 900 + 272 = \$166,172 \end{aligned}$$

This is the last interval for evaluation and yields the lowest total cost; hence, it is chosen. Where the previous total cost function for fixed holding cost was a segmented curve with offsets, this is a combination of different curves, each valid over a specific range. In the case of the fixed holding cost, it had only one minimum point, yet the offsets in the total cost due to changes in applicable unit cost in an interval could change the overall minimum. In this situation, there are different values of \hat{Q} for each unit price. The question becomes whether the value of \hat{Q} falls within the price domain. If it does, the total cost function is evaluated at that point; if not, a determination must be made as to whether the value of \hat{Q} lies to the left or right of the range. If it is to the left, the smallest value in the range is used to determine the minimum cost in the range. If it is to the right, the maximum value in the range is used.

Conclusions Regarding Inventory Models

The discussion here has covered only a small percentage of the class of deterministic inventory models, although these models represent a large percentage of applications. Should the models discussed here not adequately represent the situation under study, further research should be directed at finding a model with improved fidelity for the situation. Other models are covered in Refs. 2 and 7.

4 AGGREGATE PLANNING—MASTER SCHEDULING

Aggregate planning is the process of determining overall production, inventory, and workforce levels that are required to meet forecasted demand over a finite time horizon while trying to minimize the associated production, inventory, and workforce costs. Inputs to the aggregate planning process are forecasted demand for the products (either aggregated or individual); outputs from aggregate planning after desegregation into the individual products are the scheduled end products for the master production schedule. The time horizon for aggregate planning normally ranges from 6 to 18 months, with a 12-month average.

The difficulties associated with aggregate planning are numerous. Product demand forecasts vary widely in their accuracy; the process of developing a suitable aggregate measure to use for measuring the value or quantity of production in a multiple product environment is not always possible; actual production does not always meet scheduled production; unexpected events, including material shortages, equipment breakdowns, and employee illness, occur. Nevertheless, some form of aggregate planning is often required because seldom is there a match between the timing and quantity for product demand versus product manufacture. How the organization should staff and produce to meet this imbalance between production and fluctuating demand is what aggregate planning is about.⁸

4.1 Alternative Strategies to Meet Demand Fluctuations

Manufacturing managers use numerous approaches to meet changes in demand patterns for both short and intermediate time horizons. Among the more common are the following:

1. Produce at a constant production rate with a constant workforce, allowing inventories to build during periods of low demand, and supplying demand from inventories during periods of high demand. This approach is used by firms with tight labor markets, but customer service may be adversely affected and levels of inventories may widely fluctuate between being excessively high to being out of stock.
2. Maintain a constant workforce, but vary production within defined limits by using overtime, scheduled idle time, and potential subcontracting of production requirements. This strategy allows for rapid reaction to small or modest changes in production when faced with similar demand changes. It is the approach generally favored by many firms, if overall costs can be kept within reasonable limits.
3. Produce to demand, letting the workforce fluctuate by hiring and firing, while trying to minimize inventory carrying costs. This approach is used by firms that typically use low-skilled labor where the availability of labor is not an issue. Employee morale and loyalty, however, will always be degraded if this strategy is followed.

4.2 Aggregate Planning Costs

Aggregate planning costs can be grouped into one or more of the following categories:

1. *Production costs.* These costs include all of those items that are directly related to or necessary in the production of the product, such as labor and material costs. Supplies, equipment, tooling, utilities, and other indirect costs are also included, generally through the addition of an overhead term. Production costs are usually divided into fixed and variable costs, depending on whether the cost is directly related to production volume.
2. *Inventory costs.* These costs include the same ordering, carrying, and shortage costs discussed in inventory models.
3. *Costs associated with workforce and production rate changes.* These costs are in addition to the regular production costs and include the additional costs incurred when new employees are hired and existing employees are fired or paid overtime premiums. They may also include costs when employees are temporarily laid off or given alternative work that underutilizes their skills, or production is subcontracted to an outside vendor.

4.3 Approaches to Aggregate Planning

Researchers and practitioners alike have been intrigued by aggregate planning problems, and numerous approaches have been developed over the decades. Although difficult to categorize, most approaches can be grouped as in Table 10.

Optimal formulations take many forms. Linear programming models are popular formulations and range from the very basic, which assume deterministic demand, a fixed workforce, and no shortages, to complicated models that use piecewise linear approximations to quadratic cost functions, variable demand, and shortages.^{9–13} The linear decision rule (LDR) technique was developed in an extensive project and is one of the few instances where the approach was implemented.^{7,14–16} Nevertheless, due to the very extensive data collection, updating, and processing requirements to develop and maintain the rules, no other implementation has been reported. Lot size models usually are either of the capacitated (fixed capacity) or uncapacitated (variable capacity) variety.^{17,18} Although a number of lot size models have been developed and refined,

Table 10 Classification of Aggregate Planning Approaches

Original	Nonoptimal
Linear programming	Search techniques
Linear decision rule	Simulation models
Lot size models	Production switching heuristics
Goal programming	Management coefficient models
Other analytical aspects	—

Source. Modified from Ref. 7.

including some limited implementation, computational complexity constrains consideration to relatively small problems.¹⁹ Goal programming models are attempts at developing more realistic formulations by including multiple goals and objectives. Essentially these models possess the same advantages and disadvantages of Linear Programming (LP) models, with the additional benefit of allowing tradeoffs among multiple objectives.^{20,21} Other optional approaches have modeled the aggregate planning problem using queueing,²² dynamic programming,^{23–25} and Lagrangian techniques.^{26,27}

Nonoptimal approaches have included the use of search techniques (ST), simulation models (SM), production switching heuristics (PSH), and management coefficient models (MCM). STs involve first the development of a simulation model that describes the system under study to develop the system's response under various operating conditions. A standard search technique is then used to find the parameter settings that maximize or minimize the desired response.^{28,29} SMs also develop a model describing the firm, which is usually run using a restricted set of schedules to see which performs best. SMs allow the development of very complex systems but computationally may be so large as to disallow exhaustive testing.³⁰

PSHs were developed to avoid frequent rescheduling of workforce sizes and production rates. For example, the production rate P in period t is determined by Hwang and Cha³¹:

$$P_t = \begin{cases} L & \text{if } F_t - I_{t-1} < N - C \\ H & \text{if } F_t - I_{t-1} > N - A \\ N & \text{if } \text{otherwise} \end{cases} \quad (30)$$

where F_t = demand forecast for period t

I_{t-1} = net inventory level (inventory on hand minus back order) at the beginning of period t

L = low-level production rate

N = normal-level production rate

H = high-level production rate

A = minimum acceptable target inventory level

C = maximum acceptable target inventory level

Although this example shows three levels of production, fewer or more levels could be specified. The fewer the levels, the less rescheduling and vice versa. However, with more levels, the technique should perform better because of its ability to better track fluctuations in demand and inventory levels. MCMs were developed by attempting to model and duplicate management's past behavior.¹¹ However, consistency in past performance is required before valid models can be developed, and it has been argued that if consistency is present, the model is not required.³²

4.4 Levels of Aggregation and Disaggregation

It should be obvious from the previous discussion that different levels of aggregation and disaggregation can be derived from use of the various models. For example, many of the linear programming formulations assume aggregate measures for multiple production and demand units such as production hours, and provide output in terms of the number of production hours that must be generated per planning period. For the multiple-product situation, therefore, a scheduler at the plant level would have to disaggregate this output into the various products by planning periods to generate the master production schedule. However, if data were available to support it, a similar, albeit more complex, model could be developed that considered the individual products in the original formulation, doing away with the necessity of disaggregation. This is not often done because of the increased complexity of the resultant model, the increased data requirements, and the increased time and difficulty in solving the formulation. Also, aggregate forecasts that are used as input to the planning process are generally more accurate than forecasts for individual products.

A major task facing the planner, therefore, is determining the level of aggregation and disaggregation required. Normally this is determined by the following:

1. The decision requirements and the level of detail required. Aggregate planning at a corporate level is usually more gross than that done at a division level.
2. The amount, form, and quality of data available to support the aggregate planning process. The better the data, the better the likelihood that more complex models can be supported. Complex aggregate models may also require less disaggregation.
3. The timing frequency and resources available to the planner. Generally, the more repetitive the planning, the simpler the approach becomes. Data and analysis requirements as well as analyst's capabilities significantly increase as the complexity of the approach increases.

4.5 Aggregate Planning Dilemma

Although aggregate planning (AP) models have been available since 1955 and many variations have been developed in the ensuing decades, few implementations of these models have been reported. Aggregate planning is still an important production planning process, but many managers are unimpressed by the modeling approach. Why is that? One answer is that AP occurs throughout the organizational structure but is done by different individuals at different levels in the organization for different purposes. For example, a major AP decision is that of plant capacity, which is a constraint on all lower-level AP decisions. Determinations of if and when new plant facilities are to be added are generally corporate decisions and are made at that level. However, input for the decision comes from both division and plant levels. Division-level decision makers may then choose between competing plant facilities in their AP process in determining which plants will produce which quantity of which products within certain time frames, with input from the individual plant facilities. Plant-level managers may aggregate plan their production facilities for capacity decisions, but then must disaggregate these into a master production schedule for their facility that is constrained by corporate and division decisions.

Most models developed to date do not explicitly recognize that aggregate planning is a hierarchical decision-making process performed on different levels by different people. Therefore, AP is not performed by one individual in the organization, as implicitly assumed by many modeling approaches, but by many people with different objectives in mind.

Other reasons that have been given for the lack of general adoption of AP models include:

1. The AP modeling approach is viewed as a top-down process, whereas many organizations operate AP as a bottom-up process.

2. The assumptions to use many of the models, such as linear cost structures, the aggregation of all production into a common measure, or all workers are equal, are too simplistic or unrealistic.
3. Data requirements are too extensive or costly to obtain and maintain.
4. Decision makers are intimidated by or unwilling to deal with the complexity of the models' formulations and required analyses.

Given this, therefore, it is not surprising that few modeling approaches have been adopted in industrial settings. Although research continues on AP, there is little to indicate any significant modeling breakthrough in the new future that will dramatically change this situation. One direction, however, is to recognize the hierarchical decision-making structure of AP, and to design modeling approaches that utilize it. These systems may be different for different organizations and will be difficult to design, but currently appear to be one approach for dealing with the complexity necessary in the aggregate planning process if a modeling approach is to be followed. For a comprehensive discussion of hierarchical planning systems, see Ref. 33.

5 MATERIALS REQUIREMENTS PLANNING

Materials requirements planning is a procedure for converting the output of the aggregate planning process, the master production schedule, into a meaningful schedule for releasing orders for component inventory items to vendors or to the production department as required to meet the delivery requirements of the master production schedule. It is used in situations where the demand for a product is irregular and highly varying as to the quantity required at a given time. In these situations the normal inventory models for quantities manufactured or purchased do not apply. Recall that those models assume a constant demand and are inappropriate for the situation where demand is unknown and highly variable. The basic difference between the independent and dependent demand systems is the manner in which the product demand is assumed to occur. For the constant demand case it is assumed that the daily demand is the same. For dependent demand a forecast of required units over a planning horizon is used. Treating the dependent demand situation differently allows the business to maintain a much lower inventory level in general than would be required for the same situation under an assumed constant demand. This is so because the average inventory level will be much less in the case where MRP is applied. With MRP, the business will procure inventory to meet high demand just in advance of the requirement and at other times maintain a much lower level of average inventory.

Basic Definitions of Terms

Available units: Units of stock that are in inventory and are not in the category of buffer or safety stock and are not otherwise committed.

Gross requirements: The quantity of material required at a particular time that does not consider any available units.

Inventory unit: A unit of any product that is maintained in inventory.

Lead time: The time requirement for the conversion of inventory units into required sub-assemblies or the time required to order and receive an inventory unit.

MRP (materials requirements planning): A method for converting the end item schedule for a finished product into schedules for the components that make up the final product.

MRP-II (manufacturing resources planning): A procedural approach to the planning of all resource requirements for the manufacturing firm.

Net requirements: The units of a requirement that must be satisfied by either purchasing or manufacturing.

Product structure tree: A diagram representing the hierarchical structure of the product. The trunk of the tree would represent the final product as assembled from the sub-assemblies, and inventory units that are represented by level one, which come from subsubassemblies, and inventory units that come from the second level and so on ad infinitum.

Scheduled receipts: Material that is scheduled to be delivered in a given time bucket of the planning horizon.

Time bucket: The smallest distinguishable time period of the planning horizon for which activities are coordinated.

5.1 Procedures and Required Inputs

The *master production schedule* is a schedule devised to meet the production requirements for a product during a given planning horizon. It is normally prepared from fixed orders in the short run and product requirements forecasts for the time past for which firm product orders are available. This master production schedule together with information regarding inventory status and the product structure tree and/or the bill of materials are used to produce a planned order schedule. An example of a master production schedule is shown in Table 11.

An *MRP schedule* is the basic document used to plan the scheduling of requirements for meeting the NTS. An example is shown in Table 12. Each horizontal section of this schedule is related to a single product, part, or subassembly from the product structure tree. The first section of the first form would be used for the parent product. The following sections of the form and required additional forms would be used for the children of this parent. This process is repeated until all parts and assemblies are listed.

To use the MRP schedule, it is necessary to complete a schedule first for the parent part. Upon completion of this level-zero schedule the *bottom line* becomes the input into the schedule for each child of the parent. This procedure is followed until such time each component, assembly, or purchased part has been scheduled for ordering or production in accordance with the time requirements and other limitations that are imposed by the problem parameters. Note that if a part is used at more than one place in the assembly or manufacture of the final product, it has only one MRP schedule, which is the sum of the requirements at the various levels. The headings of the MRP schedule are as follows:

Item code. The company-assigned designation of the part or subassembly as shown on the product structure tree or the bill of materials.

Level code. The level of the product structure tree at which the item is introduced into the process. The completed product is designated level 0, subassemblies or parts that go together

Table 11 Master Production Schedule for Given Product

Part Number	Quantity Needed	Due Date
A000	25	3
A000	30	5
A000	30	8
A000	30	10
A000	40	12
A000	40	15

Table 12 Example MRP Schedule Format

Item Code	Level Code	Lot Size	Lead	On Hand	Safety Stock	Allocated																
			Time (weeks)				1	2	3	4	5	6	7	8	9	10	11	12				
							Gross requirements															
							Scheduled receipts															
							Available															
							Net requirements															
							Planned order receipts															
							Planned order releases															
							Gross requirements															
							Scheduled receipts															
							Available															
							Net requirements															
							Planned order receipts															
							Planned order releases															
							Gross requirements															
							Scheduled receipts															
							Available															
							Net requirements															
							Planned order receipts															
							Planned order releases															
							Gross requirements															
							Scheduled receipts															
							Available															
							Net requirements															
							Planned order receipts															
							Planned order releases															

to make up the completed product are level 1, subsubassemblies and parts that make up level 1 subassemblies are level 2, etc.

Lot size. The size of the lot that is purchased when an order is placed. This quantity may be an economic order quantity or a lot-for-lot purchase. (This later expression is used for a purchase quantity equal to the number required and no more.)

Lead time. The time required to receive an order from the time the order is placed. This order may be placed internally for manufacturing or externally for purchase.

On hand. The total of all units of stock in inventory.

Safety stock. Stock on hand that is set aside to meet emergency requirements.

Allocated stock. Stock on hand that has been previously allocated for use such as for repair parts for customer parts orders.

The rows related to a specific item code are designated as follows:

Gross requirements. The unit requirements for the specific item code in the specific time bucket, which is obtained from the master production schedule for the level code 0 items. For item codes at levels other than level code 0, the gross requirements are obtained from the planned order releases for the parent item. Where an item is used at more than one level in the product, its gross requirements would be the summation of the planned order releases of the items containing the required part.

Scheduled receipts. This quantity is defined at the beginning of the planning process for products that are on order at that time. Subsequently, it is not used.

Available. Those units of a given item code that are not safety stock and are not dedicated for other uses.

Net requirements. For a given item code this is the difference between gross requirements and the quantity available.

Planned order receipts. An order quantity sufficient to meet the net requirements, which are determined by comparing the net requirements to the lot size (ordering quantity) for the specific item code. If the net requirements are less than the ordering quantity, an order of the size shown as the lot size will be placed; if the lot size is LFL (lot-for-lot), a quantity equal to the net requirements will be placed.

Planned order releases. This row provides for the release of the order discussed in planned order receipts to be released in the proper time bucket such that it will arrive appropriately to meet the need of its associated planned order receipt. Note also that this planned order release provides the input information for the requirements of those item codes that are the children of this unit in subsequent generations if such generations exist in the product structure.

Example 8 Material Requirement Planning (see Ref. 1). To offer realistic problems, consider the following simple product. If you were a cub scout, you may remember building and racing a little wooden race car. Such cars come 10 in a box. Each box has 10 preformed wood blocks, 40 wheels, 40 nails for axles, and a sheet of 10 vehicle number stickers. The problem is the manufacture and boxing of these race car kits. An assembly explosion and manufacturing tree are given in Figs. 6 and 7.

Studying the tree indicates four operations. The first is to cut 50 rough car bodies from a piece of lumber. The second is to plane and slot each car body. The third is to bag 40 nails and wheels. The fourth is to box materials for 10 race cars.

The information from the production structure tree for the model car together with available information regarding lot sizes, lead time, and stock on hand is posted in the MRP schedule format to provide information for analysis of the problem. In the problem, no safety stock was prescribed and no stock was allocated for other use. This information allowed the input into the MRP format of all information shown below for the eight item codes of the product. The single input into the right side of the problem format is the MPS for the parent product, A000. With this information each of the values of the MRP schedule can be calculated. Note that the output (planned order releases) of the level 0 product multiplied by the requirements per parent unit (as shown in parentheses at the top right corner of the *child* component, in the product structure tree) becomes the *gross requirements* for the (or each) *child* of the parent part.

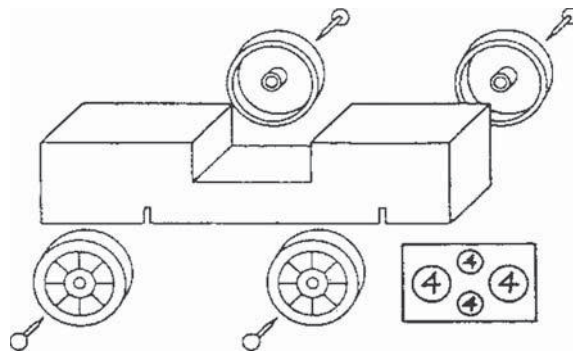


Figure 6 Diagram for model car indicating all parts.²

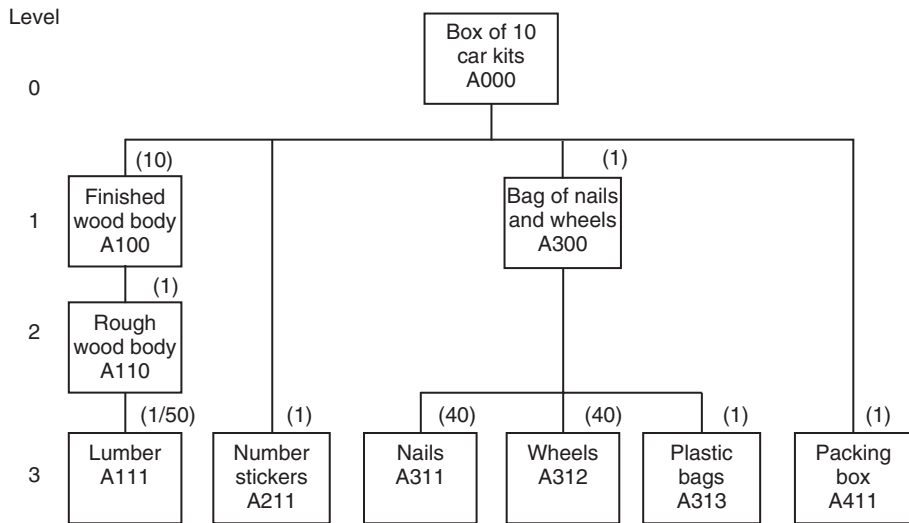


Figure 7 Product structure tree.

Calculations. As previously stated, the *gross requirements* come from either the MPS (for the parent part) or the calculation of the *planned order releases* for the parent part *times* the per-unit requirement of the current *child*, per parent part. The *scheduled receipts* are receipts scheduled from a previous MRP plan. The *available* units are those on hand from a previous period plus the *scheduled receipts* from previous MRP. The *net requirements* are *gross requirements* less the *available* units. If this quantity is negative, indicating that there is more than enough, it is set to *zero*. If it is positive, it is necessary to include an order in a previous period of quantity equal to or greater than the lot size, sufficient to meet the current need. This is accomplished by backing up a number of periods equal to the lead time for the component and placing an order in the *planned order releases* now that is equal to or greater than the lot size for the given component.

Scheduled receipts and *planned order receipts* are essentially the arrival of product. The distinction between the two is that scheduled receipts are orders that were made on a previous MRP plan. The *planned order receipts* are those that are scheduled on the current plan. Further, to keep the system operating smoothly, the MRP plan must be reworked as soon as new information becomes available regarding demand for the product for which the MPS is prepared. This essentially, provides an ability to respond and to keep materials in the *pipeline* for delivery. Without updating, the system becomes cumbersome and unresponsive. For example, most of the component parts are exhausted at the end of the 15-week period; hence, to respond in the sixteenth week would require considerable delay if the schedule were not updated. The results of this process are shown in Tables 13, 14, and 15.

The planned order release schedule (Table 16) is the result of the MRP procedure. It is essentially the summation of the bottom lines for the individual components from the MRP schedules. It displays an overall requirement for meeting the original master production schedule.

Table 13

Item Code	Level	Lot Size	Lead Time (weeks)	On Hand	Safety Stock	Allocated	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
									25	30	30	30	30	30	30	30	30	40	40			
A000	0	50	1	20	0	0	20	20	—	45	—	15	15	15	35	35	5	5	15	15	15	15
									5	—	—	—	—	15	—	—	—	35	35	25	25	25
									50	—	—	—	50	50	—	—	50	50	50	50	50	50
									500	—	—	—	500	500	—	—	500	500	500	500	500	500
A100	1	50	1	100	0	0	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
									400	—	—	—	500	500	—	—	500	500	500	500	500	500
									400	—	—	—	500	500	—	—	500	500	500	500	500	500
									50	—	—	—	50	50	—	—	50	50	50	50	50	50
A300	1	50	1	150	0	0	150	150	150	—	—	—	100	100	50	50	50	50	0	0	50	50
									0	—	—	—	100	100	50	50	50	50	0	0	50	50
									400	—	—	—	500	500	—	—	500	500	500	500	500	500
									400	—	—	—	500	500	—	—	500	500	500	500	500	500
A110	2	100	1	200	0	0	200	200	200 ^a	—	—	—	500	500	—	—	500	500	500	500	500	500
									0	—	—	—	500	500	—	—	500	500	500	500	500	500
									0	—	—	—	500	500	—	—	500	500	500	500	500	500
									0	—	—	—	500	500	—	—	500	500	500	500	500	500

^aOrdered on a previous schedule.

Table 14

Item Code	Level	Lot Size	Lead Time (weeks)	On Hand	Safety Stock	Allocated	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
							Gross requirements															
A111	3	10	3	5	0	0	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
							Scheduled receipts															
							Available															
							Net requirements															
							Planned order receipts															
							Planned order releases															
							Gross requirements															
							Scheduled receipts															
							Available															
							Net requirements															
							Planned order receipts															
							Planned order releases															
							Gross requirements															
							Scheduled receipts															
							Available															
							Net requirements															
							Planned order receipts															
							Planned order releases															
							Gross requirements															
							Scheduled receipts															
							Available															
							Net requirements															
							Planned order receipts															
							Planned order releases															
							Gross requirements															
							Scheduled receipts															
							Available															
							Net requirements															
							Planned order receipts															
							Planned order releases															

Table 15

Item Code	Level Code	Lot Size	Lead Time (weeks)	On Hand	Safety Stock	Allocated	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
							Gross requirements																
A313	3	500	3	30	0	0	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	480	
																						20	
																						500	
A411	3	500	5	40	0	0	40	40	490	490	490	490	490	440	440	440	440	440	440	390	390	390	340

^aOrdered on a previous schedule.

Table 16 Planned Ordered Release Schedule Week

	Week														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A000		50					50				50				50
A100	400					500				500		500			
A300															50
A110				500				500				500			
A111		10				10		10							
A211															
A311											2000				
A312											2000				
A313											500				
A411															

Note: An advance order of 200 units of item 110 would have to have been made on a previous MRP schedule.

5.2 Lot Sizing Techniques

Several techniques are applicable to the determination of the lot size for the order. If there are many products and some components are used in several products, it may be that demand for that common component is relatively constant. If that is the case, EOQ models such as those used in the topic on inventory can be applied.

The periodic order quantity (POQ) is a variant of the EOQ where a nonconstant demand over a planning horizon is averaged. This average is then assumed to be the constant demand. Using this value of demand, the EOQ is calculated. The EOQ is divided into the total demand if demand is greater than EOQ. This resultant figure gives the number of inventory cycles for the planning horizon. The actual forecast is then related to the number of inventory cycles and the order sizes are determined.

Example 9 Periodic Order Quantity. Given the data in Table 17 for a product that is purchased, assume that holding cost is \$10 per unit year. Order cost is \$25. Calculate the POQ. No shortage is permitted.

Using the basic EOQ formula,

$$\hat{Q} = \sqrt{\frac{2C_p D}{C_n}} = \sqrt{\frac{2(\$25)(29 \times 52)}{\$10}} = 86.9 \approx 87 \text{ units/order}$$

$$N = \frac{\text{Demand (units per 12 weeks)}}{\text{Units per order}} = \frac{348}{87} = 4 \text{ orders per 12 weeks}$$

Lot for lot is the approach to the variable demand situation, which merely requires that an order size equal to the required number of products be placed. The first order would be $25 + 29 + 34 = 88$ units. The second would be $26 + 24 + 32 = 82$ units. The third and fourth orders would be 81 and 97, respectively.

Table 17

Week	1	2	3	4	5	6	7	8	9	10	11	12	\bar{D}
Demand	25	29	34	26	24	32	28	25	28	35	32	30	29

It is coincidental that the number of orders turned out to be an integer. Had a noninteger occurred, it could be rounded to the nearest integer. An economic evaluation can be made if costs are significant, of which rounding (up or down) would yield the lower cost option. Other methods exist in the area of lot sizing.

5.3 Beyond Materials Requirements Planning

The desired output of MRP systems is the creation of a master schedule showing the time-phased production of all parts and components as well as the timing of ordering of all raw materials and components necessary to meet the forecasted product demand. Unfortunately, MRP systems do not have capacity planning capabilities to ensure that generated schedules are feasible. What evolved to remedy this deficiency was the development of manufacturing resource planning systems (referred to as MRP-II, or closed-loop systems), which make it possible to consider production capacity and supply constraints in generating integrated, feasible production schedules. Further enhancements often include shop floor control systems, which can track and monitor the execution of the production schedules and develop more effective and cost-efficient production systems. More recently (see Rondeau and Littrai³⁴), these additions have included manufacturing execution systems (MES), which are interfaces between shop floor control systems and the MRP-II system to provide not only control but some optimization capabilities with respect to the use of equipment and resultant schedules.

The growth in the features provided by these MRP-II systems with or without enhancements, however, does come at a price, in both actual dollars and added complexity. There is also the problem that while these systems are very powerful, the flexibility and capability to react to changes in the production environment because of changes in the product, improvements in materials, or stability of vendors and customers is less than that required for many modern, agile production environments. Global competition likewise has demanded that manufacturing firms be flexible and capable of quickly modifying product specifications and processes. Thus, the same or similar products may be produced in multiple facilities and used by customers from different cultures with different expectations.

To be successful in this environment, the most commonly found management outlook today appears to be customer oriented and directed toward what is known as supply chain management (SCM). SCM is used to describe the inherent linkages that exist among all of the activities of a manufacturing organization and systems and procedures that can be used to align all units of the organization so that it can be managed for overall common objectives.

A tool used in managing supply chains is enterprise resource planning (ERP). This software system is designed to integrate and optimize a variety of business processes across the entire manufacturing firm, which may include multiple sites, products, and facilities. This includes activities such as managing human resources, financial and accounting functions, sales and distribution, and inventory and manufacturing.³⁵ ERP evolved from MRP-II and generally contains three key features that go beyond MRP-II. First, ERP includes a variety of business functions, such as purchasing, sales, manufacturing, distribution, and accounting. Second, the functions are integrated so that information or data used by one activity is linked to any other activities that use that same data. More importantly, when data are entered or changed in one of the functions, it also is changed in any other function that may need or use that data. Third, each of the functions is modular in nature and can be implemented in any combination with any other business function.³⁵ This modularity and integration provides a stable information technology platform, which allows for relatively easy expansion, contraction, changes, and updates in information requirements as the business itself changes. None of this, of course, is simple, inexpensive, or easy to implement. Commitment to implement an ERP system is a strategic business decision and is done only after extensive study with decisions being made at the highest level in the organization.³⁶

6 JOB SEQUENCING AND SCHEDULING

Sequencing and scheduling problems are among the most common situations found in service and manufacturing facilities. Determining the order and deciding when activities or tasks should be done are part of the normal functions and responsibilities of management and, increasingly, of the employees themselves. These terms are often used interchangeably, but it is important to note the difference. *Sequencing* is determining the order of a set of activities to be performed, whereas *scheduling* also includes determining the specific times when each activity will be done. Thus, scheduling includes sequencing, i.e., to be able to develop a schedule for a set of activities you must also know the sequence in which those activities are to be completed.³⁷

6.1 Structure of the General Sequencing Problem

The job sequencing problem is usually stated as follows: Given n jobs to be processed on n machines, each job having a *setup time*, *processing time*, and *due date* for the completion of the job, and requiring processing on one or more of the machines, determine the sequence for processing the jobs on the machines to optimize the *performance criterion*. The factors used to describe a sequencing problem are:

1. The number of machines in the shop, m
2. The number of jobs, n
3. The type of shop or facility, i.e., job shop or flow shop
4. The manner in which jobs arrive at the shop, i.e., static or dynamic
5. The performance criterion used to measure the performance of the shop

Usual assumptions for the sequencing problem include:

1. Setup times for the jobs on each machine are independent of sequence and can be included in the processing times.
2. All jobs are available at time zero to begin processing.
3. All setup times, processing times, and due dates are known and are deterministic.
4. Once a job begins processing on a machine, it will not be preempted by another job on that machine.
5. Machines are continuously available for processing, i.e., no breakdowns occur.

Commonly used performance criteria include the following:

1. Mean flow time (F)—the average time a set of jobs spends in the shop, which includes processing and waiting times.
2. Mean idle time of machines—the average idle time for the set of machines in the shop.
3. Mean lateness of jobs (L)—the difference between the actual completion time (C_j) for a job and its due date (d_j), i.e., $L_j = C_j - d_j$. A negative value means that the job is completed early. Therefore,

$$\bar{L} = \sum_{j=1}^n \frac{C_j - d_j}{n} \quad (31)$$

4. Mean tardiness of jobs (T)—the maximum of 0 or its value of lateness, i.e., $T_j = \max\{0, L_j\}$. Therefore,

$$\bar{T} = \sum_{j=1}^n \frac{\max\{0, L_j\}}{n} \quad (32)$$

5. Mean number of jobs late.
6. Percentage of jobs late.
7. Mean number of jobs in the system.
8. Variance of lateness (S_L^2)—for a set of jobs and a given sequence, the variance calculated for the corresponding L_j 's; i.e.,

$$\sum_{j=1}^n \frac{(L_j - \bar{L})^2}{n-1} \quad (33)$$

The following material covers the broad range of sequencing problems from the simple to the complex. The discussion begins with the single machine problem and progresses through multiple machines. It includes quantitative and heuristic results for both flow shop and job shop environments.

6.2 Single-Machine Problem

In many instances the single-machine sequencing problem is still a viable problem. For example, if one were trying to maximize production through a bottleneck operation, consideration of the bottleneck as a single machine might be a reasonable assumption. For the single-machine problem, i.e., n jobs/one machine, results include the following.

Mean Flow Time

To minimize the mean flow time, jobs should be sequenced so that they are in increasing shortest processing time (SPT) order, that is,

$$t_{[1]} \leq t_{[2]} \leq \cdots \leq t_{[n]} \quad (34)$$

Example 10 Shortest Processing Time (SPT). Given are the following jobs and processing times (t_j 's) for the jobs (see Table 18).

If the jobs are processed in the shortest processing time order, i.e., (4, 2, 1, 3), then the completion times are given in Table 19. Therefore, $\bar{F} = 60/4 = 15$ days. Any other sequence will only increase \bar{F} . Proof of this is available in refs. 1, 37, and 38.

Table 18

Job j	t_j (days)
1	7
2	6
3	8
4	5

Table 19

Job j	T_j (days)	C_j (days)
4	5	5
2	6	11
1	7	18
3	8	26
$\sum C_j =$		60

Mean Lateness

As a result of the definition of lateness, SPT sequencing will minimize mean lateness (L) in the single-machine shop.

Weighted Mean Flow Time

The above results assumed that all jobs were of equal importance. What if, however, jobs should be weighted according to some measure of importance. Some jobs may be more important because of customer priority or profitability. If this importance can be measured by a weight assigned to each job, a weighted mean flow time measure, F_w , can be defined as

$$\bar{F}_w = \frac{\sum_{j=1}^n w_j F_j}{\sum_{j=1}^n w_j} \quad (35)$$

To minimize weighted mean flow time ($-F_w$), jobs should be sequenced in increasing order of weighted shortest processing time, i.e.,

$$\frac{t_{[1]}}{w_{[1]}} \leq \frac{t_{[2]}}{w_{[2]}} \leq \dots \leq \frac{t_{[n]}}{w_{[n]}} \quad (36)$$

where the brackets indicate the first, second, etc., jobs in sequence.

Example 11 Weighted Shortest Processing. Consider the data in Table 20. If jobs 2 and 6 are considered three times as important as the rest of the job, what sequence should be selected?

Solution

Job	1	2	3	4	5	6
w_j	1	3	1	1	1	3
T_j/w_j	20	9	16	6	15	8

Therefore, the job processing sequence should be 4, 6, 2, 5, 3, and 1.

Maximum Lateness/Maximum Tardiness

Other elementary results given without proof or example include the following. To minimize the maximum job lateness (L_{\max}) or the maximum job tardiness (T_{\max}) for a set of jobs, the jobs should be sequenced in order of nondecreasing due dates, i.e.,

$$d_{[1]} \leq d_{[2]} \leq \dots \leq d_{[n]} \quad (37)$$

Minimize the Number of Tardy Jobs

If the sequence above, known as the earliest due date sequence, results in zero or one tardy job, then it is also an optional sequence for the number of tardy jobs, N_T . In general, however, to

Table 20

Job	1	2	3	4	5	6
t_j (days)	20	27	16	6	15	24

find an optional sequence minimizing N_T , an algorithm attributed to Hodgson and Moore³⁹ can be used. The algorithm divides all jobs into two sets:

1. Set E , where all the jobs are either early or on time
2. Set T , where all the jobs are tardy

The optional sequence then consists of set E jobs followed by set T jobs. The algorithm is as follows:

- Step 1.** Begin by placing all jobs in set E in nondecreasing due date order, i.e., earliest due date order. Note that set T is empty.
- Step 2.** If no jobs in set E are tardy, stop; the sequence in set E is optional. Otherwise, identify the first tardy job in set E , labeling this job k .
- Step 3.** Find the job with the longest processing time among the first k jobs in sequence in set E . Remove this job from set E and place it in set T . Revise the job completion times of the jobs remaining in set E and go back to step 2 above.

Example 12 Tardy Jobs. Consider the data in Table 21.

Solution.

Step 1. $E = \{3, 1, 4, 2\}; T = \{\Phi\}$.

Step 2. Job 4 is first late job.

Step 3. Job 1 is removed from E :

$$E = \{3, 4, 2\}; T = \{1\}$$

Step 2. Job 2 is first late job.

Step 3. Job 2 is removed from

$$E = \{3, 4\}; T = \{1, 2\}$$

Step 2. No jobs in E are now late.

Therefore, optional sequences are (3, 4, 1, 2) and (3, 4, 2, 1).

6.3 Flow Shops

General flow shops can be depicted as in Fig. 8. All products being produced through these systems flow in the same direction without backtracking. For example, in a four-machine general flow shop, product 1 may require processing on machines 1, 2, 3, and 4; product 2 may require machines 1, 3, and 4; while product 3 may require machines 1 and 2 only. Thus, a flow shop processes jobs much like a production line, but, because it often processes jobs in batches, it may look more like a job shop.

Table 21

Job	t_j (days)	d_j (days)
1	10	14
2	18	27
3	2	4
4	6	16

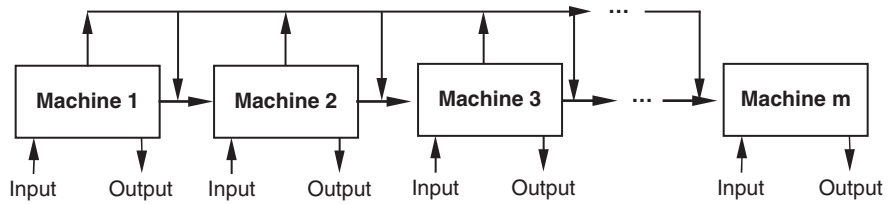


Figure 8 Product flow in a general flow shop.

Two Machines and n Jobs

The most famous result in sequencing literature is concerned with two machine flow shops and is known as *Johnson's sequencing algorithm*.⁴⁰ This algorithm will develop an optional sequence using makespan as the performance criterion. *Makespan* is defined as the time required to complete the set of jobs through all machines.

Steps for the algorithm are as follows:

1. List all processing times for the job set for machines 1 and 2.
2. Find the minimum processing time for all jobs.
3. If the minimum processing time is on machine 1, place the job first or as early as possible in the sequence. If it is on machine 2, place that job last or as late as possible in the sequence. Remove that job for further consideration.
4. Continue, by going back to step 2, until all jobs have been sequenced.

As an example, consider the five-job problem in Table 22. Applying the algorithm will give an optional sequence of (2, 4, 5, 3, 1) through the two machines, with a makespan of 26 time units.

Three Machines and n Jobs

Johnson's sequencing algorithm can be extended to a three-machine flow shop and may generate an optional solution with makespan as the criterion.

The extension consists of creating a two-machine flow shop from the three machines by summing the processing times for all jobs for the first two machines for artificial machine 1, and, likewise, summing the processing times for all jobs for the last two machines for artificial machine 2. Johnson's sequencing algorithm is then used on the two artificial machine flow shop problem.

Example 13 Flow Shop Sequencing (Johnson's Algorithm). Consider the three-machine flow shop problem in Table 23. Forming the five jobs, two artificial machine problem gives the results in Table 24.

Table 22

Job	1	2	3	4	5
Machine 1 (t_{1j})	5	1	8	2	7
Machine 2 (t_{2j})	3	4	5	6	6

Table 23

Job	Processing Times		
	t_{1j}	t_{2j}	t_{3j}
1	1	3	8
2	4	1	3
3	1	2	3
4	7	2	7
5	6	1	5

Table 24

Job	t_{aj}^1	T_{bj}^1
1	4	11
2	5	4
3	3	5
4	9	9
5	7	6

Therefore, the sequence using Johnson's sequencing algorithm is (3, 1, 4, 5, 2). It has been shown that the sequence obtained using this extension is optimal with respect to makespan if any of the following conditions hold:

1. $\min t_{1j} > \max t_{2j}$.
2. $\min t_{3j} > \max t_{2j}$.
3. If the sequence using $\{t_{1j}, t_{2j}\}$ only, i.e., the first two machines, is the same sequence as that using only $\{t_{2j}, t_{3j}\}$, i.e., only the last two machines, as two, two-machine flow shops.

The reader should check to see that the sequence obtained above is optimal using these conditions.

More Than Three Machines

Once the number of machines exceeds three, there are few ways to find optimal sequences in a flow shop environment. Enumeration procedures, like branch and bound, are generally the only practical approach that has been successfully used, and then only in problems with five or less machines. The more usual approach is to develop heuristic procedures or to use assignment rules such as priority dispatching rules. See Section 6.5 on heuristics/priority dispatching rules for more details.

6.4 Job Shops

General job shops can be represented as in Fig. 9. Products being produced in these systems may begin with any machine or process, followed by a succession of processing operations on any other sequence of machines. It is the most flexible form of production, but experience has shown that it is also the most difficult to control and to operate efficiently.

Two Machines and n Jobs

Johnson's sequencing algorithm can also be extended to a two-machine job shop to generate optimal schedules when makespan is the criterion.

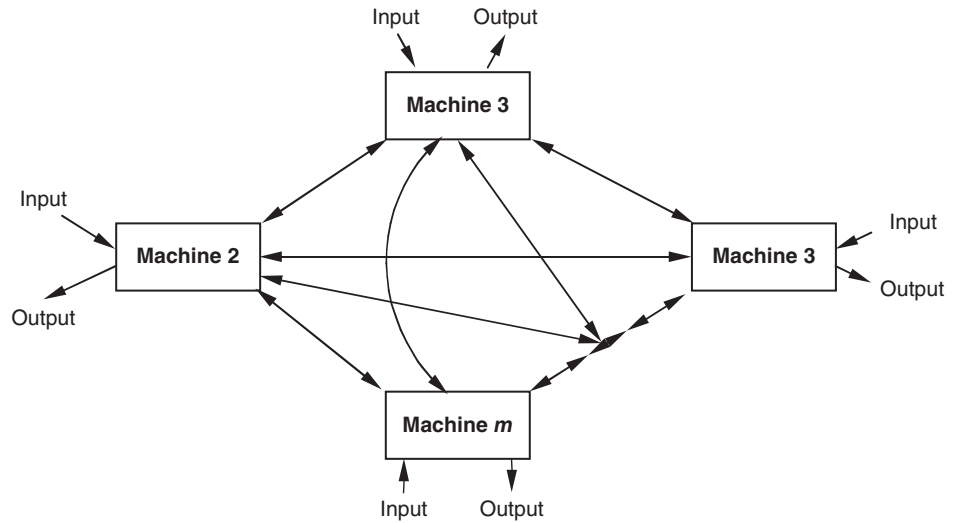


Figure 9 General job shop.

The steps to do this are as follows:

Step 1. Divide the job set into four sets:

Set {A}—jobs that require only one processing operation and that on machine 1.

Set {B}—jobs that require only one processing operation and that on machine 2.

Set {AB}—jobs that require two processing operations, the first on machine 1, the second on machine 2.

Set {BA}—jobs that require two processing operations, the first on machine 2, the second on machine 1.

Step 2. Sequence jobs in set {AB} and set {BA} using Johnson's sequencing algorithm (note that in set {BA}, machine 2 is the first machine in the process).

Step 3. The optional sequence with respect to makespan is on machine 1. Process the {AB} jobs first, then the {A} jobs, then the {BA} jobs (note that the {A} jobs can be sequenced in any order within the set). On machine 2, process the {BA} jobs first, then the {B} jobs, then the {AB} jobs (note that the {B} jobs can be sequenced in any order within the set).

For example, from Table 25 the optimal sequences are

Machine 1: 6, 5, 4, 1, 8, 9, 7, 10

Machine 2: 8, 9, 7, 10, 2, 3, 6, 5, 4

giving a makespan of 42 time units.

Machines/n Jobs

Once the problem size exceeds two machines in a job shop, optimal sequences are difficult to develop even with makespan as the criterion. If optimal sequences are desired, the only options are usually enumeration techniques like branch and bound, which attempt to take account of the special structure that may exist in the particular problem. However, because of the complexity

Table 25

Job Set	Processing Times		
	Job	t_{1j}	t_{2j}
{A}	1	3	—
	2	—	2
{B}	3	—	4
	4	4	2
{AB}	5	6	5
	6	3	7
	7	3	8
{BA}	8	4	1
	9	7	9
	10	2	4

involved in these larger problems, sequencing attention generally turns away from seeking the development of optimal schedules to the development of feasible schedules through the use of heuristic decision rules called priority dispatching rules.

6.5 Heuristics/Priority Dispatching Rules

A large number of these rules have evolved, each with their proponents, given certain shop conditions and desired performance criteria. Some of the more commonly found ones are

FCFS—select the job on a first come—first served basis

SPT—select the job with the shortest processing time

EDD—select the job with the earliest due date

STOP—select the job with the smallest ratio of remaining slack time to the number of remaining operations

LWKR—select the job with the least amount of work remaining to be done

Rules such as these are often referred to as either local or global rules. A local rule is one that is applied from the perspective of each machine or processing operation, whereas a global view is applied from the perspective of the overall shop. For example, SPT is a local rule, since deciding which of the available jobs to process next is determined by each machine or process operator. On the other hand, LWKR is global rule, since it considers all remaining processing that must be done on the job. Therefore, LWKR can be considered the global equivalent of SPT. Some rules, like FCFS can be used in either local or global applications. The choice of local or global use is often a matter of whether shop scheduling is done in a centralized or decentralized manner and whether the information system will support centralized scheduling. Implicit within these concepts is the fact that centralized scheduling requires more information to be distributed to individual workstations and is inherently a more complex scheduling environment requiring more supervisory oversight. Global scheduling intuitively should produce better system schedules, but empirical evidence seems to indicate that local rules are generally more effective.

Whichever rule may be selected, priority assignments are used to resolve conflicts. As an example of this consider the three-machine, four-job sequencing problem in Table 26. If we assume that all jobs are available at time zero, the initial job loading is shown in Fig. 10. As shown, there is no conflict on machines 1 and 2, so the first operation for jobs 1 and 2 would

Table 26

Job	Processing Times (Days)			Operation Sequence		
	t_{1j}	t_{2j}	t_{3j}	M/C-1	M/C-2	M/C-3
1	4	6	8	1	2	3
2	2	3	4	2	1	3
3	4	2	1	2	3	1
4	3	3	2	3	2	1

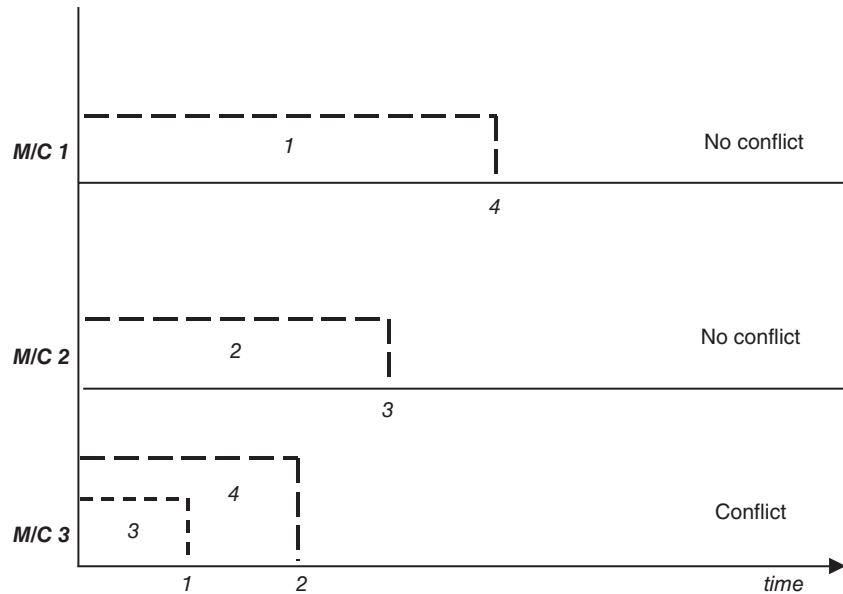


Figure 10 Stage 1 problem.

be assigned to these machines. However, there is a conflict on machine 3. If SPT were being used, the first operation for job 3 would be assigned on machine 3, and the earliest that the first operation of job 4 could be assigned to machine 3 is at time equal to 2 days.

Continuing this example following these assignments, the situation shown in Fig. 11 would then exist. If we continue to use SPT, we would assign the second operation of job 2 to machine 1 to resolve the conflict. (*Note:* If FCFS were being used, the second operation of job 3 would have been assigned.) Even though there is no conflict at this stage on machine 2, no job would normally be assigned at this time because it would introduce idle time unnecessarily within the schedule. The first operation for job 4, however, would be assigned to machine 3.

With these assignments, the schedule now appears in Fig. 12. The assignments made at this stage would include:

1. Second operation, job 4 to machine 2
2. Third operation, job 2 to machine 3, since no other job could be processed on machine, 3 during the idle time from 3 to 6 days

Note that the second operation for job 3 may or may not be scheduled at this time because the third operation for job 4 would also be available to begin processing on machine 1 at the

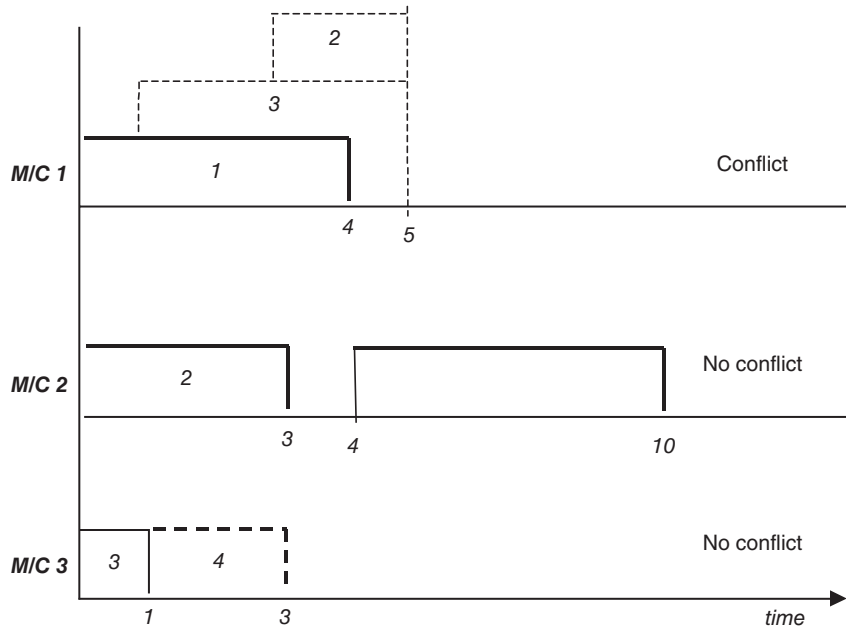


Figure 11 Stage 2 problem.

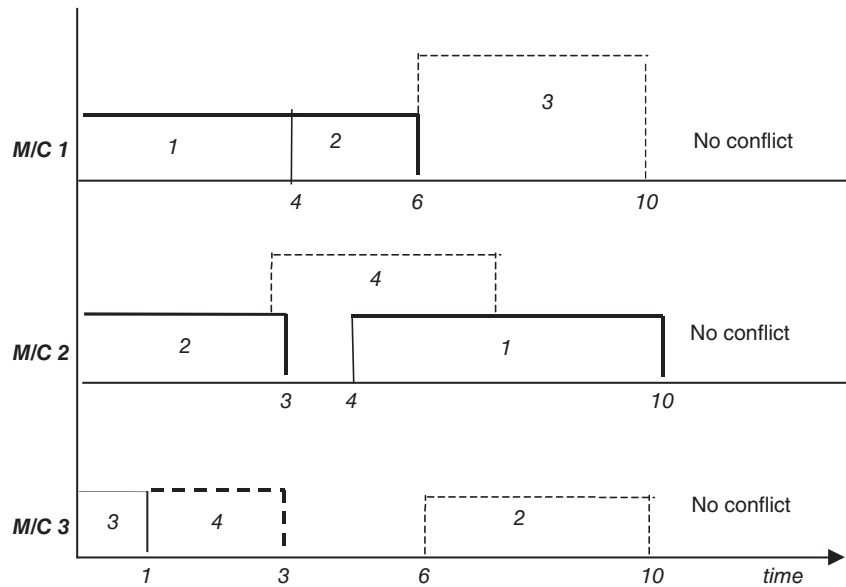


Figure 12 Stage 3 problem.

sixth day. Because of this, there would be a conflict at the beginning of the sixth day, and if SPT is being used, the third operation for job 4 would be selected over the second operation for job 3.

Several observations from this partial example can be made:

1. Depending on the order in which conflicts are resolved, two people using the same priority dispatching rule may develop different schedules.
2. Developing detailed schedules is a complex process. Almost all large-scale scheduling environments would benefit from the use of computer aids.
3. Determining the effectiveness of a dispatching rule is difficult in the schedule generating process because of the precedence relationships that must be maintained in processing.

Testing the effectiveness of dispatching rules is most often done by means of simulation studies. Such studies are used to establish the conditions found in the shop of interest for testing various sequencing strategies that management may believe to be worth investigating. For a good historical development of these as well as general conclusions that have attempted to be drawn see Ref. 41.

Two broad classifications of priority dispatching rules seemed to have emerged:

1. Those trying to reduce the flow time in which a job spends in the system, i.e., by increasing the speed going through the shop or reducing the waiting time
2. Those due-date-based rules, which may also manifest themselves as trying to reduce the variation associated with the selected performance measure

Although simulation has proven to be effective in evaluating the effectiveness of dispatching rules in a particular environment, few general conclusions have been drawn. When maximum throughput or speed is the primary criterion, SPT is often a good rule to use, even in situations when the quality of information is poor concerning due dates and processing times. When due date rules are of interest, selection is much more difficult. Results have been developed showing that when shop loads are heavy, SPT still may do well; when shop loads are moderate, STOP was preferable. Other research has shown that the manners in which due dates are set as well as the tightness of the due dates can greatly affect the performance of the rule. Overall, the following conclusions can be drawn:

1. It is generally more difficult to select an effective due-date-based rule than a flow-time-based rule.
2. If time and resources are available, the best course of action is to develop a valid model of the particular shop of interest, and experiment with the various candidate rules to determine which are most effective, given that situation.

6.6 Assembly Line Balancing

Assembly lines are viewed as one of the purest forms of production lines. A usual form is visualized as shown in Figs. 13 and 14, where work moves continuously by means of a powered conveyor through a series of workstations where the assigned work is performed.

Definitions

Cycle time (C): The time available for a workstation to perform its assigned work, assumed to be the same for each workstation. The cycle time must be greater than or equal to the longest work element for the product. Note that it is also the time between successive completions of units of product on the line.

Balance delay of a workstation: The difference between the cycle time (C) and the station time (S_j) for a workstation, i.e., the idle time for the station ($C - S_j$).

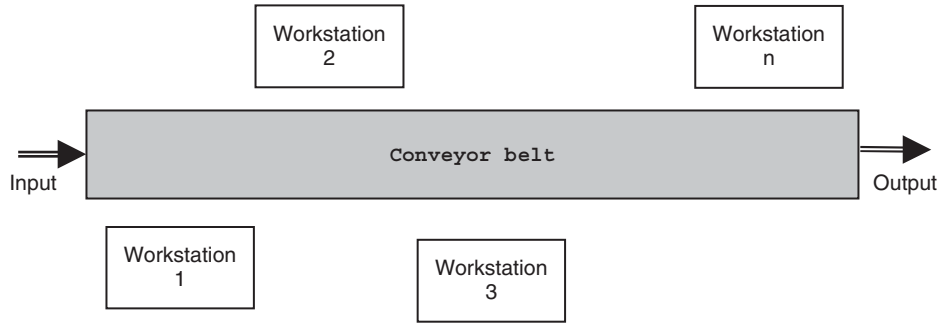


Figure 13 Structure of assembly line (conveyor type).

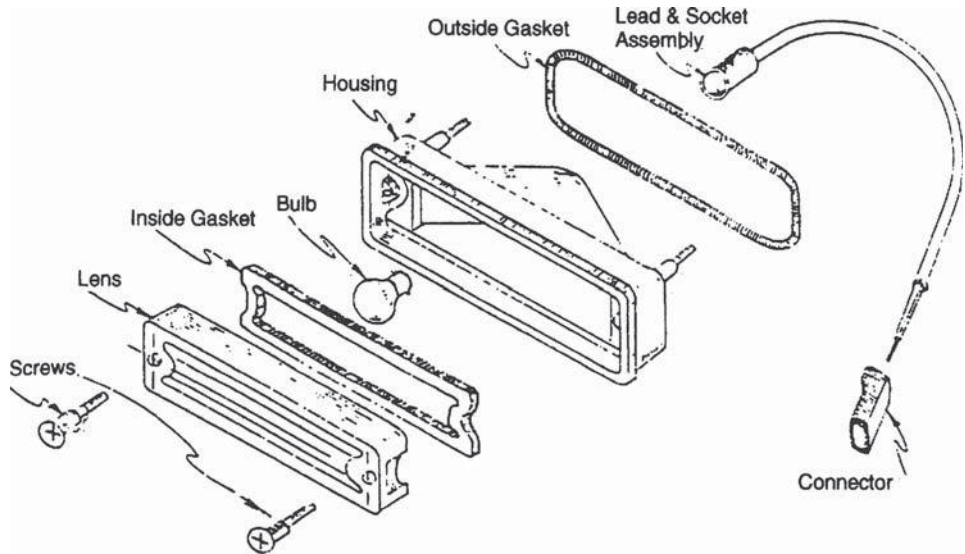


Figure 14 Automobile backup-light assembly.⁴⁰

Station time (S_j): The total amount of work assigned to station j , which consists of one or more of the work elements necessary for completion of the product. Note that each S_j must be less than or equal to C .

Work element (i): An amount of work necessary in the completion of a unit of product ($i = 1, 2, \dots, I$). It is usually considered indivisible. I is the total number of work elements necessary to complete one unit of product.

Work element time (t_i): The amount of time required to complete work element i . Therefore, the sum of all of the work elements, i.e., the total work content,

$$T = \sum_{i=1}^I t_i$$

is the time necessary to complete one unit of product.

Workstation (j): A location on the line where assigned work elements on the product are performed ($1 \leq j \leq J$).

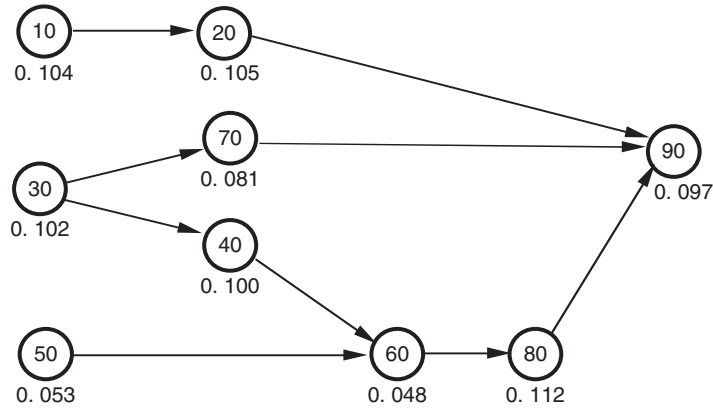


Figure 15 Precedence diagram for automobile backup-light assembly.⁴⁰

Structure of the Assembly Line Balancing Problem

The objective of assembly line balancing is to assign work elements to the workstations so as to minimize the total balance delay (total idle time) on the line. The problem is normally presented by means of a listing of the work elements and a precedence diagram, which show the relationships that must be maintained in the assembling of the product. See Figs. 14 and 15 for work content data and precedence relationship. In designing the assembly line, therefore, the work elements must be assigned to the workstations while adhering to these precedence relationships.

Note that if the balance delay is summed over the entire production line, the total balance delay (total idle time) is equal to

$$\sum_{j=1}^J (C - S_j)$$

So, to minimize the total balance delay is the same as

$$\begin{aligned} \text{Min} \sum_{j=1}^J (C - S_j) &= JC - \sum_{j=1}^J S_j \\ &= JC - (\text{Total work content for 1 unit of product}) \\ &= JC - \text{a constant} \end{aligned} \quad (38)$$

Therefore, minimizing the total balance delay is equivalent to:

1. Keeping the number of workstations constant and minimizing the cycle time, or
2. Keeping the cycle time constant and minimizing the number of workstations, or
3. Jointly trying to minimize the product of cycle time and number of workstations

Which approach might be followed could depend on the circumstances. For example, if production space was constrained, approach 1 above might be used to estimate the volume the line would be capable of producing. Approach 2 might be used if the primary concern was ensuring a certain volume of product could be produced in a certain quantity of time. Approach 3 could be used in developing alternative assignments by trading off faster line speed (shorter cycle times, more workstations, and greater production) for slower line speeds (fewer workstations, longer cycle times, and less production).

Designing the Assembly Line

Given the above structure and definitions, the following must hold:

1. $\max t_j \leq C \leq T$.
2. Minimum number of workstations = $[T/C]$, where the brackets indicate the value is rounded to the next largest integer.
3. C_{\max} = Product time available/Product volume required.

(C_{\max} is the maximum value the cycle time can be if the line is to generate the specified quantity in the specified time.) As an example, consider the data provided in Table 27 and Figs. 14 and 15.

Designing a line to produce 2000 units in a 7 1/2-hour shift would give

From condition 3:

$$C = \frac{7.5 \text{ h/shift (60 min/h)}}{2000 \text{ units(shift)}} = 0.225 \text{ min/unit}$$

From condition 2:

$$\text{Minimum number of workstations} = \frac{0.802}{0.225} = [3.56] = 4$$

Also note that condition 1 is satisfied, i.e., $0.112 \leq 0.225 \leq 0.802$.

Line Balancing Techniques

Efforts have been made to optimally model variations of these problems, but currently no procedures exist that guarantee optimal solutions to these types of problems. Practitioners and researchers, therefore, have developed a variety of heuristic procedures.⁴³ A general approach in making the assignment of work elements to workstations is to select a cycle time and to start assigning work elements where precedence restrictions are satisfied to the first workstation. Combinations of work elements may be explored to reduce the idle time present to the lowest level possible, before going to the next workstation and repeating the procedure. This process is continued until all work elements have been assigned.

Example 14 Line Balancing. Applying this procedure to the data in Table 27 and Fig. 14 would give the solution for a cycle time of 0.225 min. Also note that the cycle time could be reduced

Table 27

Workstation	Work Elements Assigned	Station Time	Balance Delay Time
1	10	0.104	0.016
	20	<u>0.105</u>	
		0.209	
2	30	0.102	0.042
	70	<u>0.081</u>	
		0.183	
3	40	0.100	0.024
	50	0.053	
	60	<u>0.048</u>	
4		0.201	0.016
	80	0.112	
	90	<u>0.097</u>	
		0.209	

to 0.209 min with this assignment and would theoretically reduce the total balance delay by $0.016 \text{ min} \times 4 \text{ workstations} = 0.064 \text{ min}$, resulting in a production increase to a total of

$$\frac{(7.5 \text{ h/shift})(60 \text{ min/shift})}{0.209 \text{ min}} = 2153 \text{ units/shift}$$

Mixed-Model Assembly Lines

The above discussion is predicated on the premise that only one product is being manufactured on the line. Many production lines are designed to produce a variety of products. Good examples of these are assembly lines that may produce several models of the same automobile with a wide variety of options. These are often referred to as mixed-model assembly lines. Similar examples are applicable in filing cabinet manufacturing, as shown in Fig. 16.^{44,45} These assembly lines are significantly more complex to design than the single-model line because of two problem areas:

1. The assignment of work elements to the workstations
2. The ordering or sequencing of models on the line

One usual approach taken in designing a mixed-model line uses the same general objective of minimizing the total balance delay (or idle time) in the assignment of work elements to workstations. However, in the mixed-model case a production period has to be defined and the assignments are made so as to minimize the total amount of idle time for all stations and models for the production period, rather than for the cycle time as in the single-model case. To use this approach the designer must define all of the work elements for all of the models and determine the quantity of each of the models being assembled within the specified production period. Once the total work content and the time allowed for production are known, work elements are assigned to workstations usually based on similarity of the work elements, tooling or equipment required, and time to perform the tasks. If the stations on the mixed-model line are not tightly linked and small in-process inventory buffers are allowed to exist between workstations, this approach seems satisfactory. However, if the stations are tightly linked where no in-process inventory is allowed between stations or if the line is operating as a just-in-time (JIT) system, this approach may not produce satisfactory assignments without analysts being especially diligent in determining the sequence of models being produced.

Determining the order of models to produce on the line is generally more difficult than the problem of assigning work elements to workstations because it has to be done in a constantly changing environment. This difficulty stems from two interrelated subproblems:

1. Trying to fully utilize the resources of the line, so that no station is idled due to a bottleneck or lack of product upon which to work

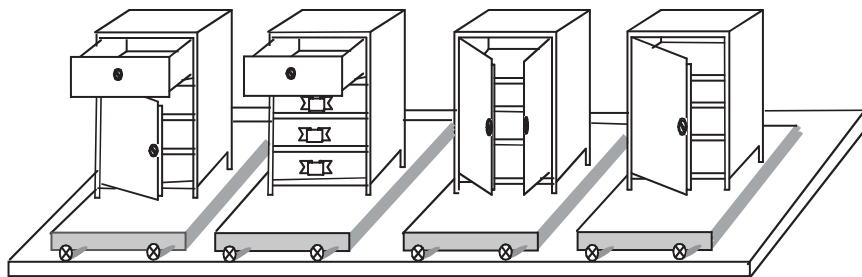


Figure 16 Filing cabinet manufacturing line (mixed model).^{44,45}

2. Trying to even out the flow of component parts to the line from upstream manufacturing or subassembly operations feeding the line, so that these resources are also fully utilized or have a relatively constant amount of work

The first of these, i.e., fully utilizing the resources of the line, is the easier of the two, especially if some flexibility is present in the system, such as producing in a make-to-stock environment or allowing small buffer, in-process inventories between workstations or variable-time workstations on the line. Examples of how some of this flexibility can be built in can be found in Ref. 46.

Smoothing out the flow of components into the assembly line is a very difficult problem, especially if the facility is operating in a JIT environment. One of the earliest approaches, which discussed how this problem was handled in the Toyota company, is known as a *goal-chasing method*.^{47,48} The procedure has since evolved into a newer version, and since it has the capability of handling multiple goals, it is called a *goals-coordinating method*.⁴⁸ This procedure has two main components: (1) appearance ratio control and (2) continuation and interval controls.

Appearance ratio control is a heuristic that determines the sequence of models on the line by attempting to minimize the variances of the components used for those products, i.e., minimize the actual variation in component usage around a calculated average usage. A production schedule of end products is built by starting with the first end product to be scheduled, then working toward the last end product. For each step in determining the sequence, the following is calculated for each product, with the minimum D determining the next product to be produced:

$$D_{Ki} = \sqrt{\sum_{j=1}^{\beta} \left(\frac{KN_j}{Q} - X_{j,K-1} - b_{ij} \right)^2} \quad (39)$$

- where
- D_{Ki} = distance to be minimized for sequence number K and for end product i
 - β = number of different components required
 - K = sequence number of the current end product in the schedule
 - N_j = total number of components j required for all products in the final schedule
 - Q = total production quantity of all end products in the final schedule
 - $X_{i,K}$ = cumulative number of components actually used through assembly sequence K
 - b_{ij} = number of components required to make one unit of end product i

However, while this approach results in a smoothed production for the majority of the schedule, it will potentially cause uneven use of components during the final phases of the day's schedule. To prevent this, continuation and interval controls are applied as constraints, which may override the appearance ratio control and introduce other type models on the line. Continuation controls ensure that no more than a designated number of consecutive end products that use a particular component are scheduled (a maximum sequencing number condition), whereas interval controls ensure that at least a designated minimum number of certain end products are scheduled between other end products that require a particular component (a minimum sequencing condition).

The overall sequencing selection process then works as follows.

- Step 1.** Appearance ration control is used to determine the first (or next) end product in the sequence.
- Step 2.** If the selected end product also satisfies the continuation and interval controls, the end product is assigned that position in the sequence. Unless all end products have been scheduled, go to step 1. Otherwise, stop, the schedule is complete.

Step 3. If the selected end product does not satisfy both the continuation and interval controls, the appearance ratio control is applied to the remaining end products, while ignoring the component that violated the continuation and/or interval controls. Out of the end products that do not require the component in question, the end product that minimizes the amount of total deviation in the following formula would be selected as the next (K th) in sequence ($j =$ component number).

$$\sum_{j=1}^n \left(\frac{\text{Total number of end product of the specified component } i}{\text{Total number of end product}} \right) \times K - \left(\begin{array}{l} \text{Accumulated number} \\ \text{of component } j \\ \text{up to } (K-1) \text{ th} \end{array} \right) + \left(\begin{array}{l} \text{Number of component} \\ j \text{ of } K\text{th additional} \\ \text{end product} \end{array} \right)$$

Unless all end products have been scheduled, go to step 1. Otherwise, stop, the schedule is complete.

As the number of models and components increases, the difficulties of developing satisfactory solutions for leveling production for mixed-model lines also increase. As this occurs the response is often to shorten the scheduled time period from, say, a day to every hour, to reduce the number of alternatives being investigated. On the one hand, this may seem desirable, particularly if the facility is operating in a JIT environment, but there is a danger that the resulting schedules will become so inefficient that they will degrade the overall performance of the line. The leveling of production on mixed-model lines remains an active research topic, with much of the research focusing on developing better or more efficient heuristic scheduling procedures.^{49,50}

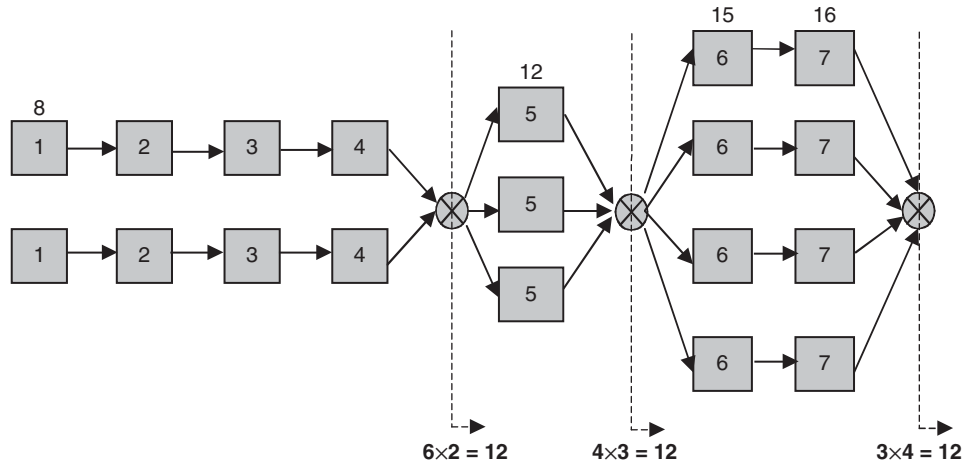
Parallel Line Balancing

When the task time of one or more elements in an assembly line becomes more than the specified cycle time, the concept of parallel line balance becomes pronounced. Helgeson and Birnie⁵¹ developed the ranked positional weight (RPW) technique and Moodie and Young⁵² developed a two-phase procedure for a serial line balancing by using a largest candidate rules and trade and transfer of elements between uneven stations. Many researchers studied the parallel line balancing problem from various scopes of the balance and techniques, but the most recent studies by Sarker and Shanthikumar⁵³ provided a more general heuristic to balance such a line applicable for both serial and parallel lines.

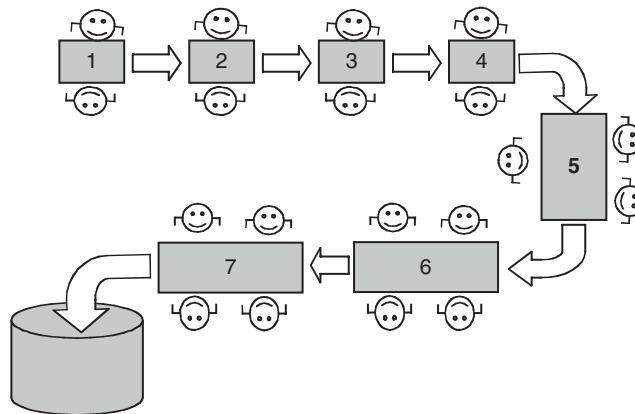
If the work element times are t_i , $i = 1, 2, \dots, n$, and C is the cycle time for the line, then the line configuration is given by the following relationship:

$$\text{Configuration} = \begin{cases} \text{Parallel line} & \text{if } \max \{t_i, i = 1, 2, \dots, n\} > C \\ \text{Serial line} & \text{if } \max \{t_i, i = 1, 2, \dots, n\} \leq C \end{cases} \quad (40)$$

Figure 17a provides a seven-stage balanced parallel line configuration for a specified cycle time of $C = 8$ min, but for an operational cycle time of $6C = 48$ min.⁵¹ The station time of the balanced line is provided at the top of each block in Fig. 17a and the number at the bottom indicates the output/48 min at the end of each line, the number of parallel lines, and the total output at the end of each parallel configuration as shown in the figure. Sarker and Shanthikumar⁵¹ provided a complete exercise of the heuristic. An equivalent serial line configuration of this parallel line configuration is depicted in Fig. 17b.



(a)



(b)

Figure 17 Parallel assembly line: (a) parallel lines and (b) linearization

7 JAPANESE MANUFACTURING PHILOSOPHY

Within the arena of manufacturing, a number of new approaches have revolutionized thinking toward designing and controlling manufacturing organizations. Foremost among this thinking have been the Japanese, who have developed and perfected a whole new philosophy. Some of these more important concepts related to production planning and control are presented below.

7.1 Just-in-Time Philosophy/Kanban Mechanism

The central concepts are to design manufacturing systems that are as simple as possible and then to design simple control procedures to control them. This does not mean that Japanese manufacturing systems are simple, but it does mean that the design is well engineered to perform the required functions and the system is neither overdesigned nor underdesigned.

Central to this philosophy is the *just-in-time (JIT)* concept. JIT is a group of beliefs and management practices that attempt to eliminate all forms of waste in a manufacturing enterprise, where waste is defined as anything not necessary in the manufacturing organization. Waste in practice may include inventories, waiting times, equipment breakdowns, scrap, defective products, and excess equipment changeover times. The elimination of waste and the resulting simplification of the manufacturing organization are the results of implementing the following related concepts usually considered as defining or making up JIT.

1. *Kanban* (the word means card) is used to control the movement and quantity of inventory through the shop, since a kanban must be attached to each container of parts. The amount of production and in-process inventory, therefore, is controlled by the number of cards that are issued to the plant floor. An additional, major benefit of using kanban is the very significant reduction in the information system that has to be used to control production.

Various forms of kanban exist, but the most frequently encountered are variations of the single-card or two-card system. One example of a two-card kanban system is that presented in Fig. 18. This example consists of two workstations, A and B. For simplicity, it is assumed that the production from workstation A is used at workstation B. The containers that move between these workstations have been sized to hold only a certain quantity of product. The two different types of kanbans used are a *withdrawal* and a *production* kanban. To control the amount of production for a given period of time, say one day, workstation B is issued a predetermined number of withdrawal kanbans. The system operates as follows:

- a. When workstation B needs parts, the operator takes an empty container, places a withdrawal kanban on it, and takes it to the storage area.
- b. The full containers in the storage area each have a production kanban on them. The worker removes the production kanban from a full container and places it on

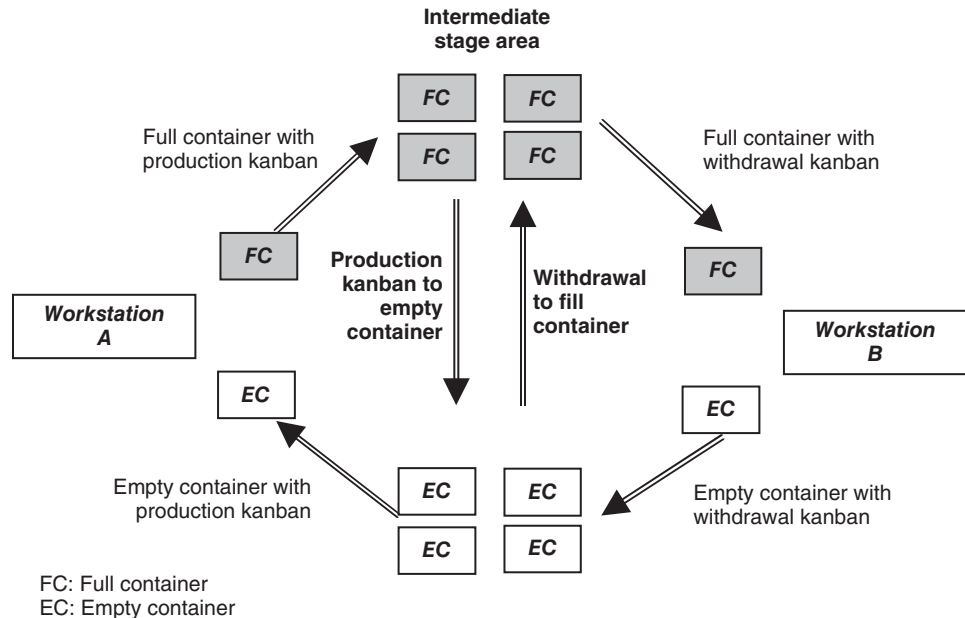


Figure 18 Two-stage kanban system.

the empty container, and removes the withdrawal kanban and places it on the full container.

- c. The worker then transports the full container (now with the withdrawal kanban) back to workstation B.
- d. Workstation A checks the production kanbans (on the empty containers) when checking for work to do. If a production kanban is present, this is the signal to begin production. If no production kanbans are present, workstation A does not continue to produce parts.

The materials in a JIT system are kept minimal at the raw materials site, the work-in-process inventory, and also the finished goods warehouse. Raw materials are procured in shipments and consumed over time more or less at a constant rate, whereas in both work-in-process and the finished goods warehouse, the products are demanded right on time instantaneously, resulting in smooth buildup followed by lumpy demand as reflected in Fig. 19. In the figure such an operation is shown with the flow of kanbans, either full or empty or partially empty.

For this system to work, certain rules have to be adhered to:

- (i) Each workstation works as long as there are *parts to work on* and a *container in which to put them*. If one or the other is missing, production stops.
 - (ii) There must be the same number of kanban cards as there are containers.
 - (iii) Containers are conveyed either full with only their standard quantities or empty.
2. *Lot size reduction* is used to reduce the amount of in-process inventory in concert with kanban, by selecting the proper size containers to use, and to increase the flexibility of the shop to change from one product to another. Overall benefits from using reduced lot sizes include shorter throughput times for product, and thus smaller lead times are required in satisfying customer orders.
 3. *Scheduling* is used to schedule small lot production to increase the flexibility of the shop in reacting to changes in demand, and to produce the quantity of goods just in time to when they are needed.
 4. *Setup time reduction* is used to reduce the times required for machines to change from one product to another so as to allow lot size reduction and JIT scheduling. Reducing

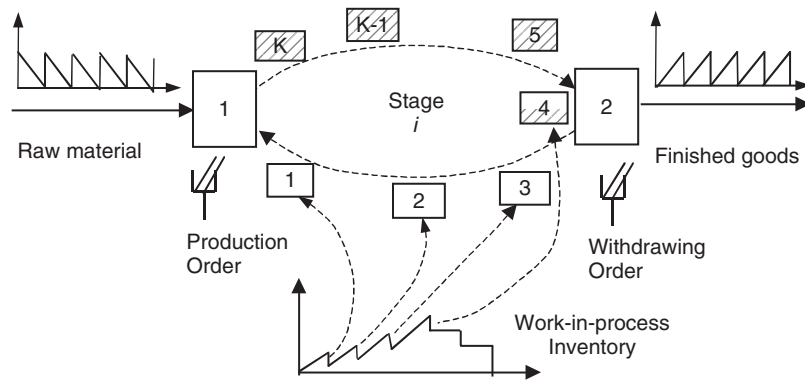


Figure 19 Single-stage JIT operational mechanics.⁵⁴

changeover times between products is critical to operating the production facility more like a flow shop and less like a job shop.

5. *Total quality management and maintenance* is used to reduce the disturbances to the manufacturing system by attempting to eliminate the making of defective products and breakdown of equipment. Central to the Japanese manufacturing philosophy is an obsession with maintenance and quality issues. For such a tightly controlled system to work it is imperative that equipment function when it is supposed to, and components and products be produced that meet or exceed customer requirements. Unexpected breakdowns or the production of bad parts is considered waste, and causes of such happenings are always high on the list for elimination in the quest for continuous improvement of the manufacturing processes.
6. *Employee cross-training* is used to provide flexibility in the workforce to allow the organization to be able to react to changes in product demand and its resultant effect on the type and quantity of employee skills required. Multiskilled workers are necessary prerequisites in any form of JIT implementation.

7.2 Time-based Competition

Following on the heels of JIT and the Japanese manufacturing philosophy is a business strategy called time-based competition. The successes of these earlier approaches were primarily grounded in providing the customer with better, more consistent quality products, which might also be less expensive in certain cases. Quality and cost were the major attributes of competitiveness for the organizations that successfully employed these techniques. Although being competitive in quality and cost will always be important, some industries are finding that this alone is not enough to maintain an edge over their competitors, since many of their competitors have also gained benefits by implementing JIT and related concepts. A third element is being introduced—that of time. Time-based competition (TBC) seeks a competitive advantage by the reduction of lead times associated with getting product to customers. TBC attempts to achieve reductions in the times required to design, manufacture, sell, and deliver products for its customers by analyzing and redesigning the processes that perform these functions.

TBC is seen as a natural evolution of JIT in that the implementation of JIT was most often found in production. Because time spent on the shop floor represents less than one-half of the time it takes to get a product to the customer for most industries, TBC is a form of extension of JIT to the rest of the manufacturing organization, including such areas as design, sales, and distribution. Wherever in the organization lead times exist that lengthen the time it takes to get the desired product to the customer, the TBC approach seeks to reduce them.

Two forms of TBC exist: first to market for new products (FM) and first to customer for existing products (FC). Companies that seek to gain a competitive advantage through FM tend to be in dynamic industries, such as those that manufacture automobiles and consumer products. For these industries, new innovations, developments, and improvements are important for their product's image and are necessary to maintain and increase product sales. Companies employing FC as a competitive advantage tend to be in more stable industries, where innovations and new product developments are less frequent and dramatic. Thus, the products that competitors sell are very similar and competitive in terms of features, price, and quality. Here the emphasis is on speed—reducing the time it takes to get the product in the customer's hands from the time at which it is ordered. There is nothing, of course, that prevents a company from employing both FM and FC approaches, and in the continuous improvement context, both approaches will be necessary if the full benefits of TBC are to be realized.

8 SUPPLY CHAIN MANAGEMENT

A concept that has gained increasing acceptance in the past 10 years is the realization that the traditional areas of procurement, production, and logistics functions that every manufacturing organization has are interrelated and should be managed as one system. The result has been the development of the term supply chain management (SCM) with further refinements and expansions so that the term now includes virtually any activity in which a manufacturing organization is involved. Common activities that go under the name of SCM include the following.

1. *Purchasing.* This may be as simple as deciding from which vendor to purchase goods and services or as complex as looking at the procurement function as a strategic business decision. Commonly embedded in purchasing when approached as an SCM component are the ideas of reducing the number of suppliers by setting up selection and certification processes, developing long-term contractual relationships that are mutually beneficial, and setting up well-defined lines of communication for supplier involvement in the organization's decisions that affect them.
2. *Production.* Activities considered SCM in internal production commonly include establishing procedures for supplier and customer involvement. Suppliers may assist in such areas as reducing quality problems associated with materials and components, developing more compatible lot sizes for production equipment, or reducing inventory requirements. Customers may assist in improving product design and quality, or suggest better packaging and shipping alternatives. Development of long-term relationships with vendors and customers may also include setting up formal committees or teams of personnel from the interested parties to work on mutually beneficial solutions to problems affecting the working relationship.
3. *Logistics.* Movement of product to customers has always been an area of great importance, but now the term is also being used to describe the process of getting materials and components to manufacturing as well as intra- and interfacilities movement. Transportation costs are one of the largest costs in making the product available to customers. Again, vendors and customers are often involved. Just-in-time scheduling of deliveries to reduce inventories for both manufacturer and customer, multiple receiving and shipping docks for receipt and shipment of items closest to point of use or convenience, receipt of items at first production process, JIT scheduling for downstream production operations, and vendor-managed inventories are some of the approaches that may be used to reduce costs and improve service.

8.1 Distribution Logistics

Fixed-interval deliveries of the manufactured goods require a reliable manufacturing system with regular supply of raw materials to ensure the production and delivery of finished goods to the customers. A scenario encountered in a production–inventory system such as a supply chain logistics system in electronics industries is depicted in Fig. 20. A silicon wafer vendor supplies wafers to company M, which manufactures Power PC chips, which, in turn, are delivered to several outside customers, such as companies A, I, and M itself. To keep the buyers' demands satisfied at different time intervals, the manufacturing company (company M in this case) has to maintain its production at a regular pace by procuring silicon wafers at regular intervals of time. Therefore, both the manufacturing company and the finished goods customers need to operate in a harmonic logistics, and to keep the wafer production–inventory system operative at minimal cost, the supply chain logistics of raw materials (silicon wafers) and finished products (Power PC chips) should be efficient (see Fig. 20).

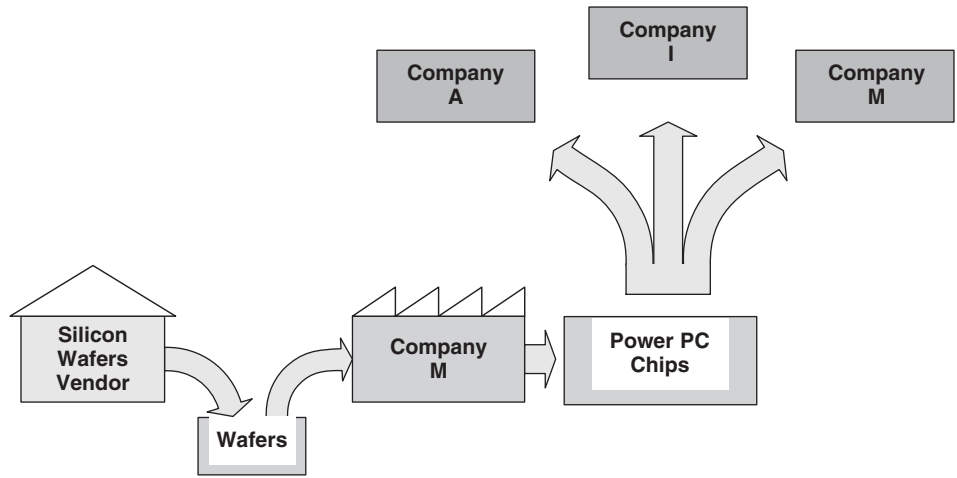


Figure 20 Supply chain logistics in electronics industries.⁵⁵

8.2 Applications of Kanban Mechanism to Supply Chain

The concept of kanban system can be applied to interplant transportation for a supply system as shown for two consecutive workstations in Fig. 21. A kanban may be conceived as a tag with a container, which is basically a truck or any other conveyance. This concept can also be extended to multistage transportation in a supply chain system as shown in Fig. 22 for a

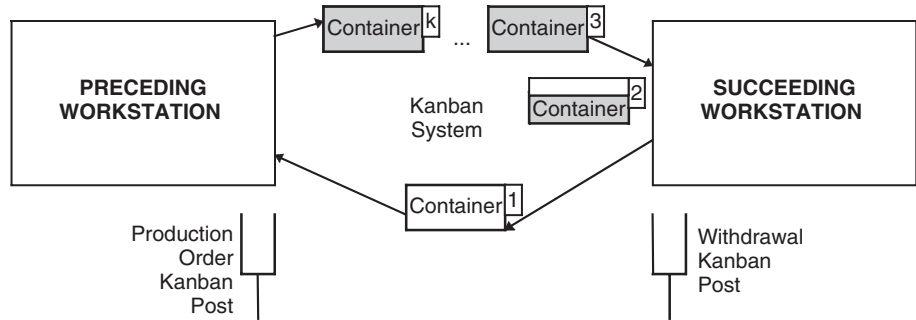


Figure 21 Single-stage kanban (SSK) system.

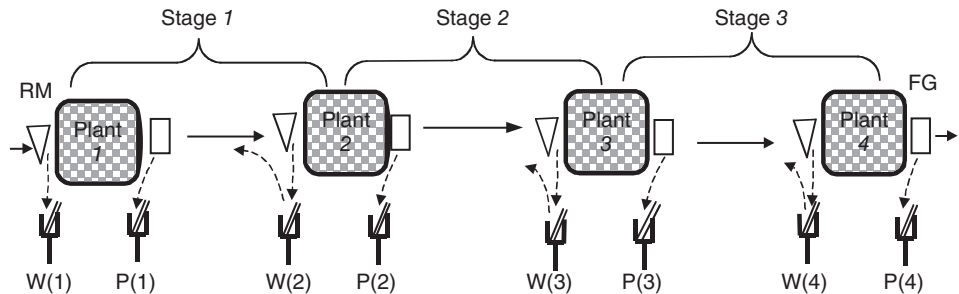


Figure 22 Three-stage supply transportation system.

three-stage supply system. The production and withdrawal kanban posts are marked with P and W followed by plant number in the parentheses.

Some of the factors for decision making in this system are itemized below:

- Transportation occurs between plants and distribution centers within the plants.
- Transporters serve as in-transit inventories that affect the MRP and inventory valuation.
- The material flow from stage to stage could be converging (assembly types) and diverging (distributive or fabrication type).
- Safety stock, inventory, and demand fluctuations play an important role in the reliability of the supply system.
- The integrated control mechanism should be synchronized for all components and parts between plants and workstations for smooth supply between any two or more points.
- Online control through modern techniques is essential for better service and customer relation. Also, an exchange of information promotes centralized versus. decentralized storage and mode of transportation.
- Transporters and their schedules.
- Transportation types and modes require decision making.
- Control of interplant supplies.
- Bulk deliveries by trucks or other intermodal mechanisms.
- Congestion problems in supply mode and warehouses.
- Delivery problems due to unscheduled shipments or dispatching.
- Lack of logistics capability analysis and optimal strategy contribute to poor performance of the system.

8.3 General Remarks

One of the underlying requirements for SCM to work is trust and commitment among the parties involved, i.e., vendors, manufacturers, and customers. As a prerequisite for successful implementation, many SCM improvements may require that significant time and resources be spent in developing these interorganizational relationships, but it is equally important that the intraorganizational relationships be developed. Getting manufacturing to work with marketing or with the financial people may represent some of the greatest challenges and obstacles. It is also imperative that effective working relationships, once established, be nourished to ensure continuing open lines of communication.

As with most management approaches applied to manufacturing, the objectives of SCM are to reduce costs and improve customer service. What is relatively new in the view of SCM is that by linking all of the business, manufacturing, and logistics functions, a competitive advantage in the marketplace may be achieved. The difficulty, of course, is that while this is conceptually feasible, in reality, the overall system is large and complex, with many subsystems whose goals are not necessarily compatible with each other. Thus, while the global goal may be total systems integration, current SCM efforts are less comprehensive and focus on opportunities where potential for improvements are the greatest, such as those mentioned above.⁵⁶ Although total systems integration is unlikely to occur in the foreseeable future, continuing to improve the linking of functions and aligning of goals of the units of the manufacturing organization can do much to sustain the continuing drive to reduce costs and improve customer service.

REFERENCES

1. D. D. Bedworth and J. E. Bailey, *Integrated Production Control Systems*, Wiley, New York, 1982.
2. E. A. Silver and R. Peterson, *Decision Systems for Inventory Management and Production Planning*, Wiley, New York, 1985.
3. L. A. Johnson, and D. C. Montgomery, *Operations Research in Production Planning: Scheduling and Inventory Control*, Wiley, New York, 1974.
4. S. Nahmias, *Production and Operations Analysis*, 5th ed., Irwin/McGraw-Hill, New York, 2005.
5. A. H. Taha, *Operations Research*, 4th ed., MacMillian, New York, 1998.
6. D. P. Gover and G. L. Thompson, *Programming and Probability Models for Operations Research*, Wadsworth, 1973.
7. S. Nam and R. Logendran, "Aggregate Production Planning—A Survey of Models and Methodologies," *Eur. J. Oper. Res.*, **61**, 255–272, 1992.
8. A. C. Hax, "Aggregate Production Planning," in *Production Handbook*, 4th ed., J. White (Ed.), Wiley, New York, 1987, pp. 3.116–3.127.
9. A. Charnes, W. W. Cooper, and Mellon, B., "A Model for Optimizing Production by Reference to Cost Surrogates," *Econometrics*, **23**, 307–323, 1955.
10. E. H. Bowman, "Production Scheduling by the Transportation Method of Linear Programming," *Oper. Res.*, **4**, 100–103, 1956.
11. E. H. Bowman, "Consistency and Optimality in Managerial Decision Making," *Manage. Sci.*, **9**, 310–321, 1963.
12. M. E. Posner and W. Szwarc, "A Transportation Type Aggregate Production Model with Backordering," *Manage. Sci.*, **29**, 188–199, 1983.
13. K. Singhal and V. Adlakha, "Cost and Shortage Trade-offs in Aggregate Production Planning," *Decision Sci.*, **20**, 158–164, 1989.
14. C. C. Holt, F. Modigliani, and H. A. Simon, "A Linear Decision Rule for Production and Employment Scheduling," *Manage. Sci.*, **2**, 1–30, 1955.
15. C. C. Holt, F. Modigliani, and J. F. Muth, "Derivation of a Linear Decision Rule for Production and Employment," *Manage. Sci.*, **2**, 159–177, 1956.
16. C. C. Holt, F. Modigliani, J. F. Muth, and H. A. Simon, *Planning Production Inventories and Work Force*, Prentice-Hall, Englewood Cliffs, NJ, 1960.
17. A. S. Manne, "Programming of Economic Lot Sizes," *Manage. Sci.*, **4**, 115–135, 1958.
18. H. M. Wagner and T. M. Whitin, "Dynamic Version of the Economic Lot Size Model," *Manage. Sci.*, **5**, 89–96, 1958.
19. S. Gorenstein, "Planning Tire Production," *Manage. Sci.*, **17**, B72–B82, 1970.
20. S. M. Lee and L. J. Moore, "A Practical Approach to Production Scheduling," *J. Product. Invent. Manage.*, **15**, 79–92, 1974.
21. R. F. Deckro and J. E. Hebert, "Goal Programming Approaches to Solving Linear Decision Rule Based Aggregate Production Planning Models," *IIE Trans.*, **16**, 308–315, 1984.
22. D. P. Gaver, "Operating Characteristics of a Simple Production, Inventory-Control Model," *Oper. Res.*, **9**, 635–649, 1961.
23. W. I. Zangwill, "A Deterministic Multiproduct, Multifacility Production and Inventory Model," *Oper. Res.*, **14**, 486–507, 1966.
24. W. I. Zangwill, "A Deterministic Multiperiod Production Scheduling Model with Backlogging," *Manage. Sci.*, **13**, 105–119, 1966.
25. W. I. Zangwill, "Production Smoothing of Economic Lot Sizes with Non-decreasing Requirements," *Manage. Sci.*, **13**, 191–209, 1966.
26. G. D. Eppen and F. J. Gould, "A Lagrangian Application to Production Models," *Oper. Res.*, **16**, 819–829, 1968.
27. D. R. Lee and D. Orr, "Further Results on Planning Horizons in the Production Smoothing Problem," *Manage. Sci.*, **23**, 490–498, 1977.
28. C. H. Jones, "Parametric Production Planning," *Manage. Sci.*, **13**, 843–866, 1967.

29. A. D. Flowers and S. E. Preston, "Work Force Scheduling with the Search Decision Rule," *OMEG*, **45**, 473–479, 1977.
30. W. B. Lee and B. M. Khumawala, "Simulation Testing of Aggregate Production Planning Models in an Implementation Methodology," *Manage. Sci.*, **20**, 903–911, 1974.
31. H. Hwang and C. N. Cha, "An Improved Version of the Production Switching Heuristic for the Aggregate Production Planning Problems," *Int. J. Product. Res.*, **33**, 2567–2577, 1995.
32. S. Eilon, "Five Approaches to Aggregate Production Planning," *AIIE Trans.*, **7**, 118–131, 1975.
33. A. C. Hax and D. Candea, *Production and Inventory Management*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
34. P. J. Rondeau and L. A. Littral, "Evolution of Manufacturing Planning and Control Systems: From Reorder Point to Enterprise Resource Planning," *Product. Invent. Manage. J.*, 2nd Quarter, 1–7, 2001.
35. V. A. Mabert, A. Soni, and M. A. Venkataramanan, "Enterprise Resource Planning: Common Myths Versus Evolving Reality," *Bus. Horizons*, May–June, 69–76, 2001.
36. A. Clewett, D. Franklin, and A. McCown, *Network Resource Planning for SAP R/3, BAAN IV, and PEOPLESOFT: A Guide to Planning Enterprise Applications*, McGraw-Hill, New York, 1998.
37. S. French, *Sequencing and Scheduling: An Introduction to the Mathematics of the Job Shop*, Ellis Horwood Ltd, Halsted Press, Chichester, England, 1982.
38. K. Baker, *Introduction to Sequencing and Scheduling*, Wiley, New York, 1974.
39. T. J. Hodgson and J. M. Moore, "A Technical Note to 'Sequencing n Jobs on One Machine to Minimize the Number of Tardy Jobs,'" *Manage. Sci.*, **17**(1), 102–109, 1968.
40. S. M. Johnson, "Optimal Two- and Three-Stage Production Schedules with Setup Times Included," *Naval Res. Logistics Quart.*, **1**(1), 1954.
41. S. S. Panwalkar and W. Iskander, "A Survey of Scheduling Rules," *Oper. Res.*, **25**, 45–61, 1977.
42. J. Lorenz and D. Pooch, "Assembly Line Balancing," in *Production Handbook*, 4th ed., J. White (Ed.), Wiley, New York, 1987, pp. 3.176–3.189.
43. E. A. Elsayed and T. Boucher, *Analysis and Control of Production Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1994.
44. B. R. Sarker and H. Pan, "Designing a Mixed-Model, Open-Station Assembly Line Using Mixed-integer Programming," *J. Oper. Res. Soc.*, **52**(5), 545–558, 2001.
45. B. R. Sarker and H. Pan, "Design Configuration for a Closed-Station, Mixed-Model Assembly Line: A Filing Cabinet Manufacturing System," *Int. J. Product. Res.*, **39**(10), 2251–2270, 2001.
46. N. Thomopoulos, "Mixed Model Line Balancing with Smoothed Station Assignments," *Manage. Sci.*, **16**(9), 593–603, 1970.
47. Y. Moden, *Toyota Production System: Practical Approach to Production Management*, Industrial Engineering and Management Press, Atlanta, GA, 1983.
48. Y. Moden, *Toyota Production System: An Integrated Approach to Just-In-Time*, 2nd ed., Industrial Engineering and Management Press, Atlanta, GA, 1993.
49. R. T. Sumichrast and R. S. Russell, "Evaluating Mixed-Model Assembly Line Sequencing Heuristics for Just-in-Time Production Systems," *J. Oper. Manage.*, **9**(3), 371–386, 1990.
50. J. F. Bard, A. Shtub, and S. B. Joshi, "Sequencing Mixed-Model Assembly Lines to Level Parts Usage and Minimize Line Length," *Int. Product. Res.*, **32**(10), 2431–2454, 1994.
51. W. B. Helgeson and D. P. Birnie, "Assembly Line Balancing Using the Ranke Positional Weight Technique," *J. Ind. Eng. (old name of IIE Trans.)*, **16**(6), 394–398, 1963.
52. C. L. Moodie and H. H. Young, "A Heuristic Method of Assembly Line Balancing," *J. Ind. Eng. (old name of IIE Trans.)*, **12**(6), 394–398, 1965.
53. B. R. Sarker and J. G. Shanthikumar, "A Generalized Approach for Serial or Parallel Line Balancing," *Int. J. Product. Res.*, **21**(1), 109–133, 1983.
54. S. Wang and B. R. Sarker, "A Single-Stage Supply Chain System Controlled by Kanbans Under Just-in-Time Philosophy," *J. Operat. Res. Soc.*, **55**(5), 485–494, 2004.
55. G. R. Parija and B. R. Sarker, "Operations Planning in a Supply Chain System with Fixed-Interval Deliveries to Multiple Customers," *IIE Trans. Special Issue on Manufacturing Logistics*, **31**(11), 1075–1082, 1999.

56. I. J. Chen and A. Paulraj, "Understanding Supply Chain Management: Critical Research and a Theoretical Framework," *Int. J. Product. Res.*, **42**(1), 131–163, 2004.

BIBLIOGRAPHY

- E. A. Adam, Jr., and R. J. Ebert, *Production and Operations Management*, 4th ed., Prentice-Hall, Englewood Cliffs, NJ, 1989.
- P. Carter, S. Melnyk, and R. Handfield, "Identifying the Basic Process Strategies for Time Based Competition," *Product. Invent. Manage. J.*, 1st Quarter, 65–70, 1995.
- S. Eilon, "Aggregate Production Scheduling," in *Handbook of Industrial Engineering*, G. Salvendy, (Ed.), Wiley Interscience, New York, 1982, pp. 11.3.1–11.3.23.
- W. J. Fabrycky, P. M. Ghare, and P. E. Torgersen, *Applied Operations Research and Management Science*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
- N. Gaither, *Production and Operations Management*, 6th ed., Dryden Press, 1992.
- L. S. Lasdon, and R. C. Tedung, "An Efficient Algorithm for Multi-item Scheduling," *Operat. Res.*, **19**, 946–966, 1971.
- J. M. Mellichamp, and R. M. Love, "Production Switching Heuristics for the Aggregate Planning Problem," *Manage. Sci.*, **24**, 1242–1251, 1978.
- V. S. Nori, and B. R. Sarker, "Cyclic Scheduling for a Multi-product, Single-Facility Production System Operating under a Just-in-Time Delivery Policy," *J. Operat. Res. Soc.*, **47**(7), 930–935, 1996.
- B. R. Sarker, and C. V. Balan, "Cell Formation with Operations Times for Even Distribution of Workloads," *Int. J. Product. Res.*, **34**(5), 1447–1468, 1996.
- R. J. Schonberger, *Japanese Manufacturing Techniques*, Free Press, New York, 1983, pp. 219–245.
- B. Scott, D. N. Burt, W. Copacino, C. Gopal, H. L. Lee, R. P. Lynch, and S. Morris, "Supply Chain Challenges: Building Relationships: A Conversation," *Harvard Bus. Rev.*, **81**(7), 64, 2003.
- T. E. Volimann, W. L. Berry, and D. C. Whybark, *Manufacturing Planning and Control Systems*, 3rd ed., Irwin, Homewood, IL, 1992.
- A. M. Wagner, *Principles of Management Science*, Prentice-Hall, New York, 1970.

CHAPTER 4

PRODUCTION PROCESSES AND EQUIPMENT

Magd E. Zohdi, William E. Biles, and Dennis B. Webster
Louisiana State University
Baton Rouge, Louisiana

1 METAL-CUTTING PRINCIPLES	116	12 SHAPING, PLANING, AND SLOTTING	149
2 MACHINING POWER AND CUTTING FORCES	119	13 SAWING, SHEARING, AND CUTTING OFF	152
3 TOOL LIFE	121	14 MACHINING PLASTICS	153
4 METAL-CUTTING ECONOMICS	123	15 GRINDING, ABRASIVE MACHINING, AND FINISHING	153
4.1 Cutting Speed for Minimum Cost (V_{min})	126	15.1 Abrasives	154
4.2 Tool Life Minimum Cost (T_m)	126	15.2 Temperature	156
4.3 Cutting Speed for Maximum Production (V_{max})	126	16 NONTRADITIONAL MACHINING	158
4.4 Tool Life for Maximum Production (T_{max})	126	16.1 Abrasive Flow Machining	158
5 CUTTING-TOOL MATERIALS	126	16.2 Abrasive Jet Machining	159
5.1 Cutting-Tool Geometry	127	16.3 Hydrodynamic Machining	159
5.2 Cutting Fluids	128	16.4 Low-Stress Grinding	159
5.3 Machinability	128	16.5 Thermally Assisted Machining	160
5.4 Cutting Speeds and Feeds	129	16.6 Electromechanical Machining	160
6 TURNING MACHINES	129	16.7 Total Form Machining	161
6.1 Lathe Size	132	16.8 Ultrasonic Machining	161
6.2 Break-Even Conditions	132	16.9 Water-Jet Machining	163
7 DRILLING MACHINES	133	16.10 Electrochemical Deburring	166
7.1 Accuracy of Drills	138	16.11 Electrochemical Discharge Grinding	166
8 MILLING PROCESSES	140	16.12 Electrochemical Grinding	167
9 GEAR MANUFACTURING	143	16.13 Electrochemical Honing	167
9.1 Machining Methods	144	16.14 Electrochemical Machining	168
9.2 Gear Finishing	146	16.15 Electrochemical Polishing	169
10 THREAD CUTTING AND FORMING	146	16.16 Electrochemical Sharpening	169
10.1 Internal Threads	146	16.17 Electrochemical Turning	170
10.2 Thread Rolling	148	16.18 Electrostream	170
11 BROACHING	148	16.19 Shaped-Tube Electrolytic Machining	171

16.20	Electron Beam Machining	172	16.28	Chemical Machining: Chemical Milling, Chemical Blanking	178
16.21	Electrical Discharge Grinding	173	16.29	Electropolishing	178
16.22	Electrical Discharge Machining	173	16.30	Photochemical Machining	178
16.23	Electrical Discharge Sawing	174	16.31	Thermochemical Machining	179
16.24	Electrical Discharge Wire Cutting (Traveling Wire)	174	16.32	Rapid Prototyping and Rapid Tooling	180
16.25	Laser Beam Machining	175	REFERENCES		181
16.26	Laser Beam Torch	176	BIBLIOGRAPHY		181
16.27	Plasma Beam Machining	177			

1 METAL-CUTTING PRINCIPLES

Material removal by chipping process began as early as 4000bc, when the Egyptians used a rotating bowstring device to drill holes in stones. Scientific work developed starting about the mid nineteenth century. The basic chip-type machining operations are shown in Fig. 1.

Figure 2 shows a two-dimensional type of cutting in which the cutting edge is perpendicular to the cut. This is known as *orthogonal* cutting, as contrasted with the three-dimensional *oblique* cutting shown in Fig. 3. The main three cutting velocities are shown in Fig. 4. The metal-cutting factors are defined as follows:

α	rake angle
β	friction angle
γ	strain
λ	chip compression ratio, t_2/t_1
μ	coefficient of friction
ψ	tool angle
τ	shear stress
ϕ	shear angle
Ω	relief angle
A_o	cross section, wt_1
e_m	machine efficiency factor
f	feed rate ipr (in./revolution), ips (in./stroke), mm/rev (mm/revolution), or mm/stroke
f_t	feed rate (in./tooth, mm/tooth) for milling and broaching
F	feed rate, in./min (mm/s)
F_c	cutting force
F_f	friction force
F_n	normal force on shear plane
F_s	shear force
F_t	thrust force
HP_c	cutting horsepower
HP_g	gross horsepower
HP_u	unit horsepower
N	revolutions per minute
Q	rate of metal removal, in. ³ /min
R	resultant force
T	tool life in minutes
t_1	depth of cut
t_2	chip thickness
V	cutting speed, ft/min
V_c	chip velocity
V_s	shear velocity

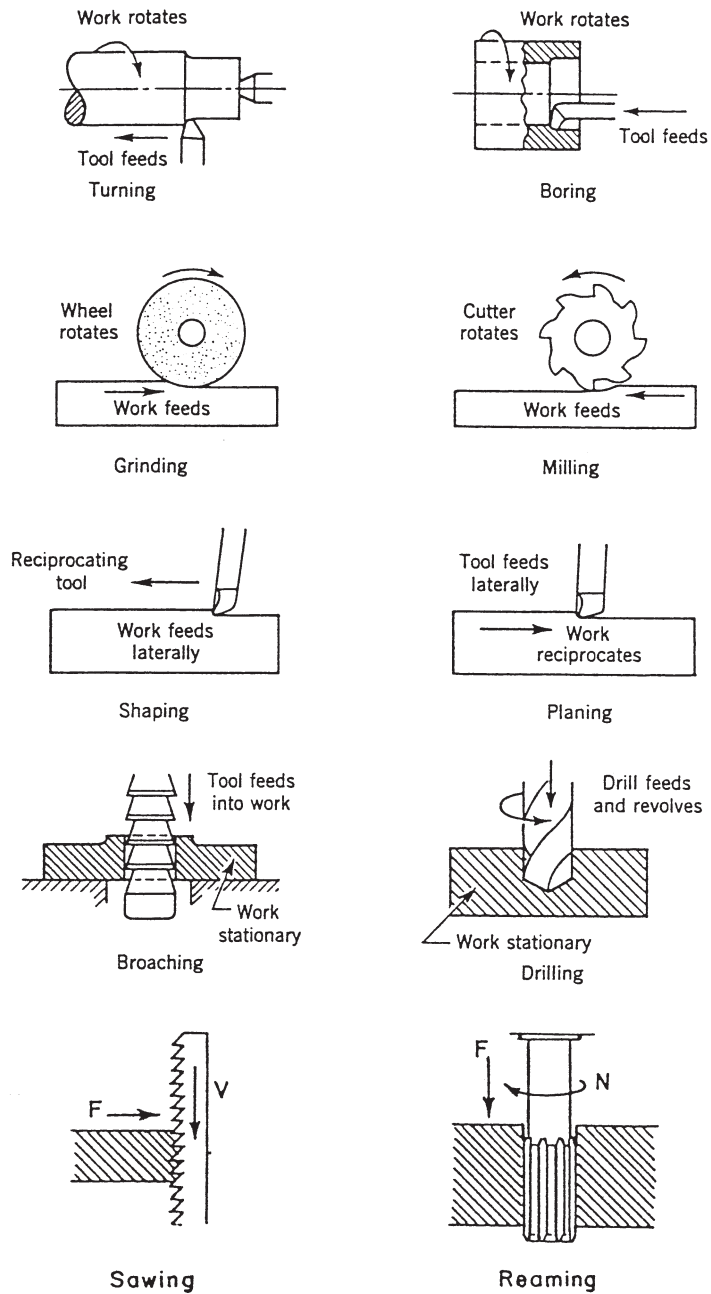


Figure 1 Conventional machining processes.

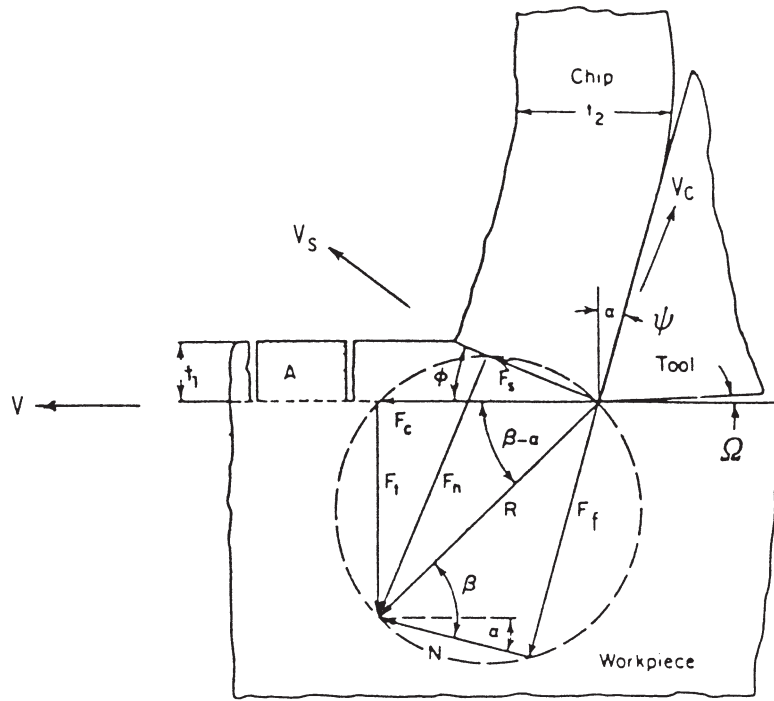


Figure 2 Mechanics of metal-cutting process.

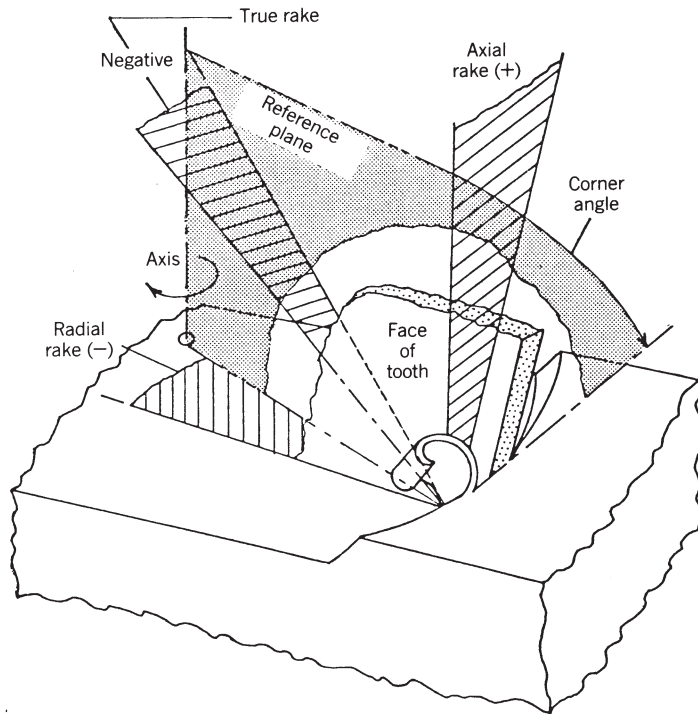


Figure 3 Oblique cutting.

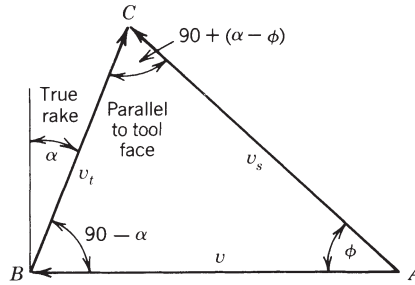


Figure 4 Cutting velocities.

The *shear angle* ϕ controls the thickness of the chip and is given by

$$\tan \phi = \frac{\cos \alpha}{\lambda - \sin \alpha} \quad (1)$$

The *strain* γ that the material undergoes in shearing is given by

$$\gamma = \cot \phi + \tan(\phi - \alpha)$$

The *coefficient of friction* μ on the face of the tool is

$$\mu = \frac{F_t + F_c \tan \alpha}{F_c - F_t \tan \alpha} \quad (2)$$

The *friction force* F_t along the tool is given by

$$F_t = F_t \cos \alpha + F_c \sin \alpha$$

Cutting forces are usually measured with dynamometers and/or wattmeters. The shear stress τ in the shear plane is

$$\tau = \frac{F_c \sin \phi \cos \phi - F_t \sin^2 \phi}{A}$$

The speed relationships are

$$\begin{aligned} \frac{V_c}{V} &= \frac{\sin \phi}{\cos(\phi - \alpha)} \\ V_c &= V/\lambda \end{aligned} \quad (3)$$

2 MACHINING POWER AND CUTTING FORCES

Estimating the power required is useful when planning machining operations, optimizing existing ones, and specifying new machines. The power consumed in cutting is given by

$$\text{Power} = F_c V \quad (4)$$

$$\text{HP}_c = \frac{F_c V}{33,000} \quad (5)$$

$$= Q \text{HP}_u \quad (6)$$

where F_c = cutting force, lb

V = cutting speed, ft/min = $\pi DN/12$ (rotating operations)

D = diameter, in.

N = revolutions/min

HP_u = specific power required to cut a material at a rate of 1 in.³/min

Q = material removal rate, in.³/min

For SI units,

$$\text{Power} = F_c V \text{ watts} \quad (7)$$

$$= QW \text{ watts} \quad (8)$$

where F_c = cutting force, newtons

$$V = \text{m/s} = 2\pi RN$$

W = specific power required to cut a material at a rate of $1 \text{ mm}^3/\text{s}$

Q = material removal rate, mm^3/s

The specific energies for different materials, using sharp tools, are given in Table 1.

$$\text{Power} = F_c V = F_c 2\pi RN$$

$$= F_c R 2\pi N$$

$$= M 2\pi N \quad (9)$$

$$= \frac{MN}{63,025} \text{ HP} \quad (10)$$

where M = torque, in.-lbf

N = revolutions per min

In SI units,

$$= \frac{MN}{9549} \text{ kW} \quad (11)$$

Table 1 Average Values of Energy per Unit Material Removal Rate

Material	BHN ^a	HP _c /in. ³ per min	W/mm ³ /s
Aluminum alloys	50–100	0.3	0.8
	100–150	0.4	1.1
Cast iron	125–190	0.5	1.6
	190–250	1.6	4.4
Carbon steels	150–200	1.1	3.0
	200–250	1.4	3.8
	250–350	1.6	4.4
Leaded steels	150–175	0.7	1.9
Alloy steels	180–250	1.6	4.4
	250–400	2.4	6.6
Stainless steels	135–275	1.5	4.1
Copper	125–140	1.0	2.7
Copper alloys	100–150	0.8	2.2
Leaded brass	60–120	0.7	1.9
Unleaded brass	50	1.0	2.7
Magnesium alloys	40–70	0.2	0.55
	70–160	0.4	1.1
Nickel alloys	100–350	2.0	5.5
Refractory alloys (tantalum, columbium, molybdenum)	210–230	2.0	5.5
Tungsten	320	3.0	8.0
Titanium alloys	250–375	1.3	3.5

^aBHN, Brinell hardness number.

where M = newton-meter
 $\text{HP/in.}^3/\text{min } 2.73 = ? \text{ W}/(\text{mm}^3/\text{s})$
 $M = F_c R = \text{power}/2\pi N$
 $F_c = M/R$

$$\text{Gross power} = e_m / \text{Cutting power} \quad (12)$$

The cutting horsepowers for different machining operations are given below.

For turning, planing, and shaping,

$$\text{HP}_c = (\text{HP}_u) 12CWVfd \quad (13)$$

For milling,

$$\text{HP}_c = (\text{HP}_u) CWFwd \quad (14)$$

For drilling,

$$\text{HP}_c = (\text{HP}_u) CW(N)f \left(\frac{\pi D^2}{4} \right) \quad (15)$$

For broaching,

$$\text{HP}_c = (\text{HP}_u) 12CWVn_c w d_t \quad (16)$$

where V = cutting speed, fpm (ft/min)

C = feed correction factor

f = feed, ipr (turning and drilling), ips (planing and shaping)

F = feed, ipm (in./min) = $f \times N$

d = depth of cut, in.

d_t = maximum depth of cut per tooth, in.

n_c = number of teeth engaged in work

w = width of cut, in.

W = tool wear factor

Specific energy is affected by changes in feed rate. Table 2 gives feed correction factor (C). Cutting speed and depth of cut have no significant effect on power. Tool wear effect factor (W) is given in Table 3. The gross power is calculated by applying the overall efficiency factor (e_m).

3 TOOL LIFE

Tool life is a measure of the length of time a tool will cut satisfactorily and may be measured in different ways. Tool wear, as in Fig. 5, is a measure of tool failure if it reaches a certain

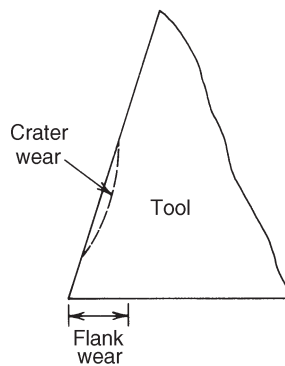
Table 2 Feed Correction (C) Factors for Turning, Milling, Drilling, Planing, and Shaping

Feed(ipr or ips)	mm/rev or mm/stroke	Factor
0.002	0.05	1.4
0.005	0.12	1.2
0.008	0.20	1.05
0.012	0.30	1.0
0.020	0.50	0.9
0.030	0.75	0.80
0.040	1.00	0.80
0.050	1.25	0.75

Table 3 Tool Wear Factors (W)

Type of Operations ^a	W
<i>Turning</i>	
Finish turning (light cuts)	1.10
Normal rough and semifinish turning	1.30
Extra-heavy-duty rough turning	1.60–2.00
<i>Milling</i>	
Slab milling	1.10
End milling	1.10
Light and medium face milling	1.10–1.25
Extra-heavy-duty face milling	1.30–1.60
<i>Drilling</i>	
Normal drilling	1.30
Drilling hard-to-machine materials and drilling with a very dull drill	1.50
<i>Broaching</i>	
Normal broaching	1.05–1.10
Heavy-duty surface broaching	1.20–1.30

^aFor all operations with sharp cutting tools.

**Figure 5** Types of tool wear.

limit. These limits are usually 0.062 in. (1.58 mm) for high-speed tools and 0.030 in. (0.76 mm) for carbide tools. In some cases, the life is determined by surface finish deterioration and an increase in cutting forces. The cutting speed is the variable that has the greatest effect on tool life. The relationship between tool life and cutting speed is given by the Taylor equation:

$$VT^n = C \quad (17)$$

where V = cutting speed, fpm (m/s)

T = tool life, min (s)

n = exponent depending on cutting condition

C = constant, the cutting speed for a tool life of 1 min

Table 4 gives the approximate ranges for the exponent n . Taylor's equation is equivalent to

$$\log V = C - n \log T \quad (18)$$

which when plotted on log–log paper gives a straight line, as shown in Fig. 6.

Table 4 Average Values of n

Tool Material	Work Material	n
HSS (18-4-1)	Steel	0.15
	C.I.	0.25
	Light metals	0.40
Cemented carbide	Steel	0.30
	C.I.	0.25
Sintered carbide	Steel	0.50
Ceramics	Steel	0.70

Equation (19) incorporates the size of cut:

$$K = VT^n f^{n_1} d^{n_2} \quad (19)$$

Average values for $n_1 = 0.5-0.8$
 $n_2 = 0.2-0.4$

Equation (20) incorporates the hardness of the workpiece:

$$K = VT^n f^{n_1} d^{n_2} (\text{BHN})^{1.25} \quad (20)$$

4 METAL-CUTTING ECONOMICS

The efficiency of machine tools increases as cutting speeds increase, but tool life is reduced. The main objective of metal-cutting economics is to achieve the optimum conditions, that is, the minimum cost while considering the principal individual costs: machining cost, tool cost, tool-changing cost, and handling cost. Figure 7 shows the relationships among these four factors.

$$\text{Machining cost} = C_o t_m \quad (21)$$

where C_o = operating cost per minute, which is equal to the machine operator's rate plus appropriate overhead

t_m = machine time in minutes, which is equal to $L/(fN)$, where L is the axial length of cut

$$\text{Tool cost per operation} = C_t \frac{t_m}{T} \quad (22)$$

where C_t = tool cost per cutting edge

T = tool life, which is equal to $(C/V)^{1/n}$

$$\text{Tool changing cost} = C_o t_c (t_m/T) \quad (23)$$

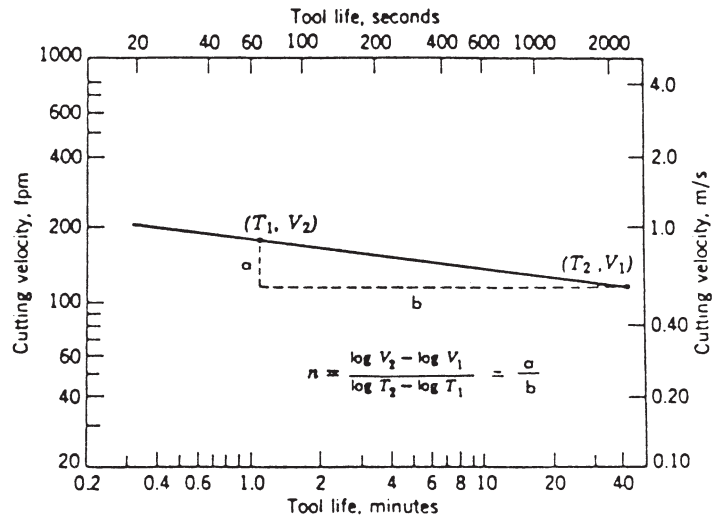
where t_c = tool changing time, min

$$\text{Handling cost} = C_o t_h \quad (24)$$

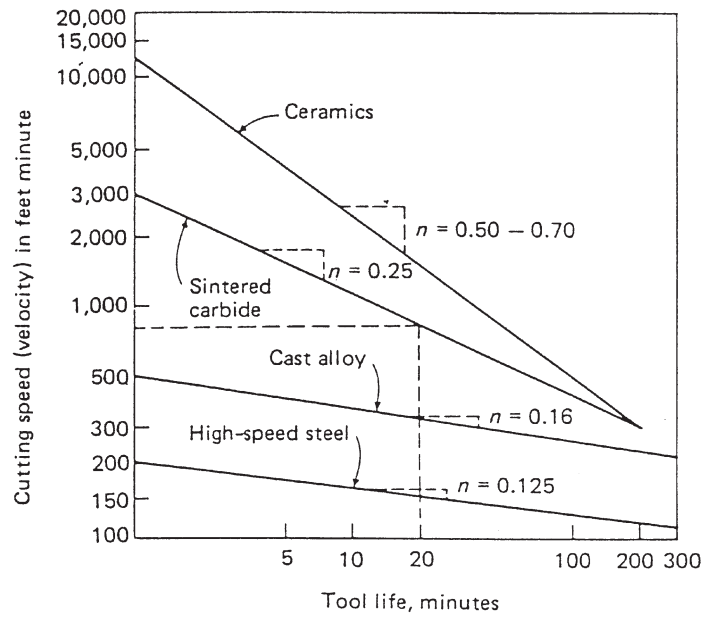
where t_h = handling time, min

The average unit cost C_u will be equal to

$$C_u = C_o t_m + \frac{t_m}{T} (C_t + C_o t_c) + C_o t_h \quad (25)$$

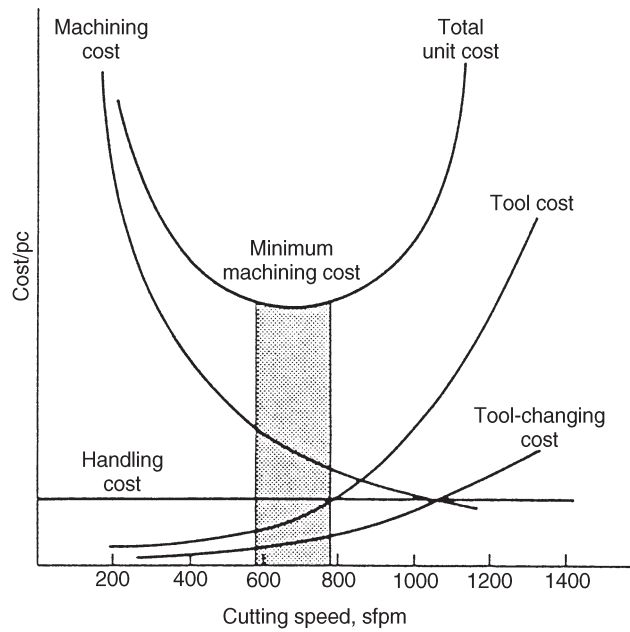


(a)

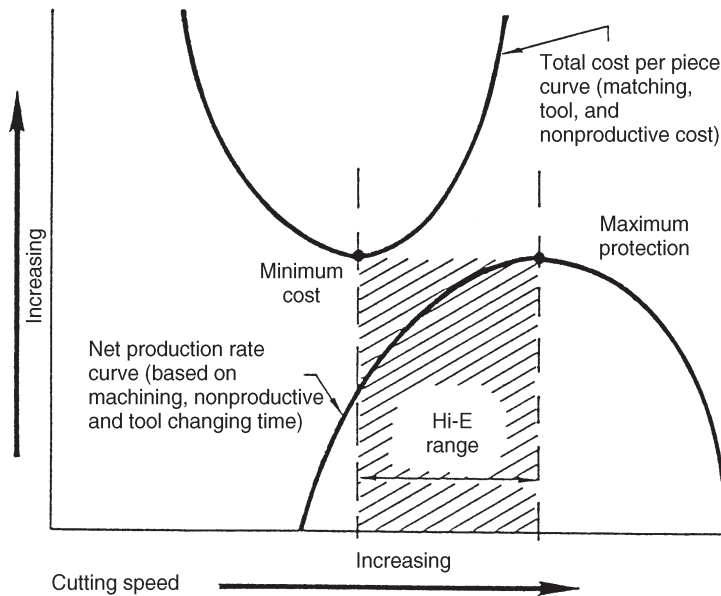


(b)

Figure 6 Cutting speed/tool life relationship.



(a)



(b)

Figure 7 Cost factors.

4.1 Cutting Speed for Minimum Cost (V_{\min})

Differentiating the costs with respect to cutting speed and setting the results equal to zero will result in V_{\min} :

$$V_{\min} = \frac{C}{(1/n - 1)[(C_o t + C_t)/C_o]^n} \quad (26)$$

4.2 Tool Life Minimum Cost (T_m)

Since the constant C is the same in Taylor's equation and Eq. (22), and if V corresponds to V_{\min} , then the tool life that corresponds to the cutting speed for minimum cost is

$$T_{\min} = \left(\frac{1}{n} - 1\right) \left(\frac{C_o t_c + C_t}{C_o}\right) \quad (27)$$

4.3 Cutting Speed for Maximum Production (V_{\max})

This speed can be determined from Eq. (26) for the cutting speed for minimum cost by assuming that the tool cost is negligible, that is, by setting $C_t = 0$:

$$V_{\max} = \frac{C}{[(1/n - 1)t_c]^n} \quad (28)$$

4.4 Tool Life for Maximum Production (T_{\max})

By analogy to Taylor's equation, the tool life that corresponds to the maximum production rate is given by

$$T_{\max} = \left(\frac{1}{n} - 1\right) t_c \quad (29)$$

5 CUTTING-TOOL MATERIALS

The desirable properties for any tool material include the ability to resist softening at high temperature, which is known as red hardness; a low coefficient of friction; wear resistance; sufficient toughness and shock resistance to avoid fracture; and inertness with respect to workpiece material.

The principal materials used for cutting tools are carbon steels, cast nonferrous alloys, carbides, ceramic tools or oxides, and diamonds.

1. *High-carbon steels* contain (0.8–1.2%) carbon. These steels have good hardening ability, and with proper heat treatment hold a sharp cutting edge where excessive abrasion and high heat are absent. Because these tools lose hardness at around 600°F (315°C), they are not suitable for high speeds and heavy-duty work.
2. *High-speed steels* (HSS) are high in alloy contents such as tungsten, chromium, vanadium, molybdenum, and cobalt. High-speed steels have excellent hardenability and will retain a keen cutting edge to temperatures around 1200°F (650°C).
3. *Cast nonferrous alloys* contain principally chromium, cobalt, and tungsten, with smaller percentages of one or more carbide-forming elements, such as tantalum, molybdenum, or boron. Cast-alloy tools can maintain good cutting edges at temperatures up to 1700°F (935°C) and can be used at twice the cutting speed as HSS and still maintain the same feed. Cast alloys are not as tough as HSS and have less shock resistance.
4. *Carbides* are made by powder-metallurgy techniques. The metal powders used are tungsten carbide (WC), cobalt (Co), titanium carbide (TiC), and tantalum carbide (TaC) in

different ratios. Carbide will maintain a keen cutting edge at temperatures over 2200°F (1210°C) and can be used at speeds two or three times those of cast alloy tools.

5. *Coated tools*, cutting tools, and inserts are coated by titanium nitride (TiN), titanium carbide (TiC), titanium carbonitride (TiCN), aluminum oxide (Al₂O₃), and diamond. Cutting speeds can be increased by 50% due to coating.
6. *Ceramic or oxide tool* inserts are made from aluminum oxide (Al₂O₃) grains with minor additions of titanium, magnesium, or chromium oxide by powder-metallurgy techniques. These inserts have an extremely high abrasion resistance and compressive strength, lack affinity for metals being cut, resistance to cratering and heat conductivity. They are harder than cemented carbides but lack impact toughness. The ceramic tool softening point is above 2000°F (1090°C), and these tools can be used at high speeds (1500–2000 ft/min) with large depth of cut. Ceramic tools have tremendous potential because they are composed of materials that are abundant in the earth's crust. Optimum cutting conditions can be achieved by applying negative rake angles (5–7°), rigid tool mountings, and rigid machine tools.
7. *Cubic boron nitride* (CBN) is the hardest material presently available, next to diamond. CBN is suitable for machining hardened ferrous and high-temperature alloys. Metal removal rates up to 20 times those of carbide cutting tools were achieved.
8. *Single-crystal diamonds* are used for light cuts at high speeds of 1000–5000 fpm to achieve good surface finish and dimensional accuracy. They are used also for hard materials difficult to cut with other tool material.
9. *Polycrystalline diamond* cutting tools consist of fine diamond crystals, natural or synthetic, that are bonded together under high pressure and temperature. They are suitable for machining nonferrous metals and nonmetallic materials.

5.1 Cutting-Tool Geometry

The shape and position of the tool relative to the workpiece have a very important effect in metal cutting. There are six single-point tool angles critical to the machining process. These can be divided into three groups.

Rake angles affect the direction of chip flow, the characteristics of chip formation, and tool life. Positive rake angles reduce the cutting forces and direct the chip flow away from the material. Negative rake angles increase cutting forces but provide greater strength, as is recommended for hard materials.

Relief angles avoid excessive friction between the tool and workpiece and allow better access of coolant to tool–work interface.

The *side cutting edge angle* allows the full load of the cut to be built up gradually. The *end cutting edge angle* allows sufficient clearance so that the surface of the tool behind the cutting point will not rub over the work surface.

The purpose of the *nose radiuses* is to give a smooth surface finish and to increase the tool life by increasing the strength of the cutting edge. The elements of the single-point tool are written in the following order: back rake angle, side rake angle, end relief angle, side relief angle, end cutting edge angle, side cutting edge angle, and nose radius. Figure 8 shows the basic tool geometry.

Cutting tools used in various machining operations often appear to be very different from the single-point tool in Figure 8. Often they have several cutting edges, as in the case of drills, broaches, saws, and milling cutters. Simple analysis will show that such tools are comprised of a number of single-point cutting edges arranged so as to cut simultaneously or sequentially.

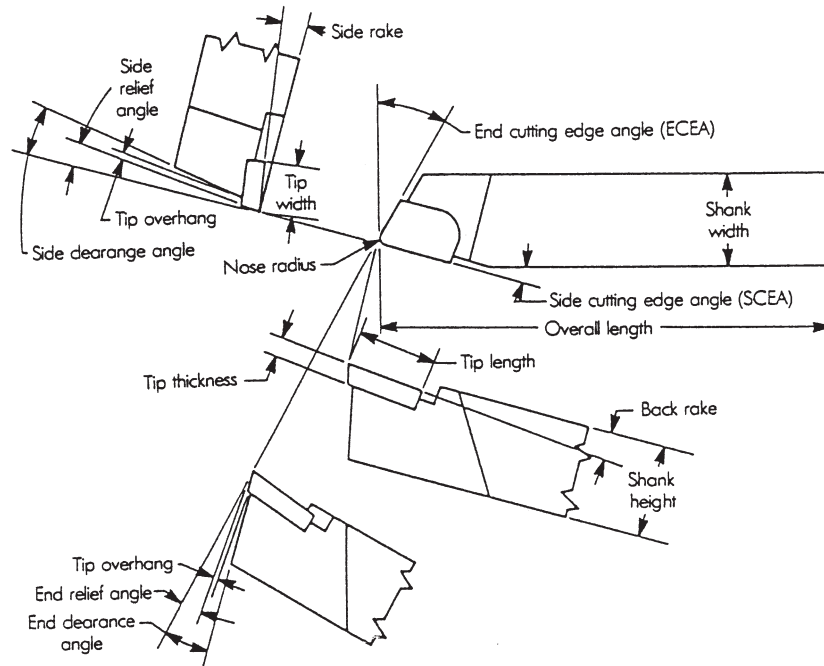


Figure 8 Basic tool geometry.

5.2 Cutting Fluids

The major roles of the cutting fluids—liquids or gases—are:

1. Removal of the heat friction and deformation
2. Reduction of friction among chip, tool, and workpiece
3. Washing away chips
4. Reduction of possible corrosion on both workpiece and machine
5. Prevention of buildup edges

Cutting fluids work as coolants and lubricants. Cutting fluids applied depend primarily on the kind of material being used and the type of operation. The four major types of cutting fluids are:

1. Soluble oil emulsions with water-to-oil ratios of 20:1 to 80:1
2. Oils
3. Chemicals and synthetics
4. Air

At low cutting speeds (40 ft/min and below), oils are highly recommended, especially in tapping, reaming, and gear and thread machining. Cutting fluids with the maximum specific heat, such as soluble oil emulsions, are recommended at high speeds.

5.3 Machinability

Machinability refers to a system for rating materials on the basis of their relative ability to be machined easily, long tool life, low cutting forces, and acceptable surface finish. Additives such

as lead, manganese sulfide, or sodium sulfide with percentages less than 3% can improve the machinability of steel and copper-based alloys, such as brass and bronze. In aluminum alloys, additions up to 1–3% of zinc and magnesium improve their machinability.

5.4 Cutting Speeds and Feeds

Cutting speed (CS) is expressed in feet per minute (m/s) and is the relative surface speed between the cutting tool and the workpiece. It may be expressed by the simple formula $CS = \pi DN/12$ fpm in., where D is the diameter of the workpiece in inches in case of turning or the diameter of the cutting tool in case of drilling, reaming, boring, and milling, and N is the revolutions per minute. If D is given in millimeters, the cutting speed is $CS = \pi DN/60,000$ m/s.

Feed refers to the rate at which a cutting tool advances along or into the surface of the workpiece. For machines in which either the workpiece or the tool turns, feed is expressed in inches per revolution (ipr) (mm/rev). For reciprocating tools or workpieces, feed is expressed in inches per stroke (ips) (mm/stroke).

The recommended cutting speeds, and depth of cut that resulted from extensive research, for different combinations of tools and materials under different cutting conditions can be found in many references, including Society of Manufacturing Engineers (SME) publications such as *Tool and Manufacturing Engineers Handbook*¹; *Machining Data Handbook*²; Metcut Research Associates, Inc.; *Journal of Manufacturing Engineers*; *Manufacturing Engineering Transactions*; *American Society for Metals (ASM) Handbook*³; *American Machinist's Handbook*⁴; *Machinery's Handbook*⁵; American Society of Mechanical Engineering (ASME) publications; Society of Automotive Engineers (SAE) publications; and *International Journal of Machine Tool Design and Research*.

6 TURNING MACHINES

Turning is a machining process for generating external surfaces of revolution by the action of a cutting tool on a rotating workpiece, usually held in a lathe. Figure 9 shows some of the external operations that can be done on a lathe. When the same action is applied to internal surfaces of revolution, the process is termed *boring*. Operations that can be performed on a lathe are turning, facing, drilling, reaming, boring, chamfering, tapping, grinding, threading, tapping, and knurling.

The primary factors involved in turning are speed, feed, depth of cut, and tool geometry. Figure 10 shows the tool geometry along with the feed (f) and depth of cut (d). The CS is the

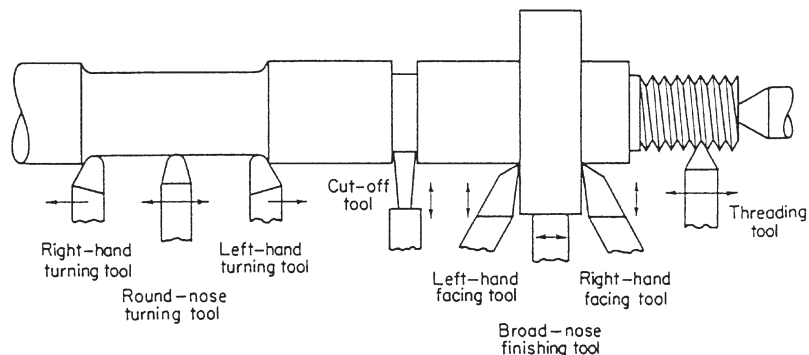


Figure 9 Common lathe operations.

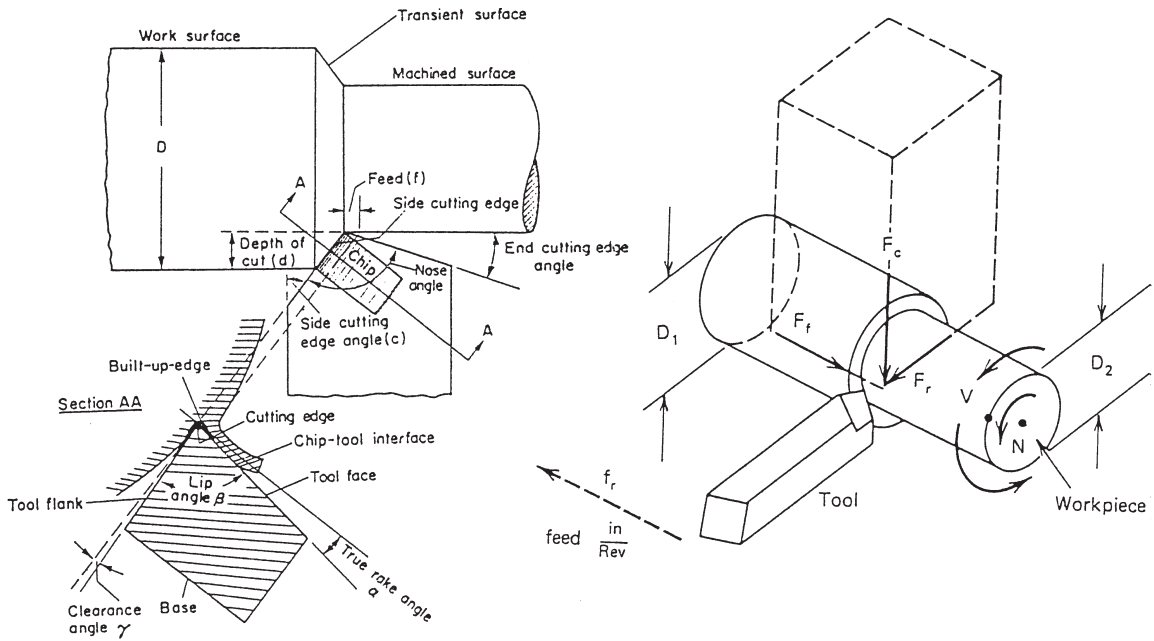


Figure 10 Tool geometry—external turning.

surface speed in feet per minute (fpm) or meters per second (m/s). The feed (f) is expressed in inches of tool advance per revolution of the spindle (ipr) or (mm/rev). The depth of cut (d) is expressed in inches. Table 5 gives some of the recommended speeds while using HSS tools and carbides for the case of finishing and rough machining. The cutting speed (fpm) is calculated by

$$CS = \frac{\pi DN}{12} \text{ fpm} \tag{30}$$

where D = workpiece diameter, in.
 N = spindle revolutions per minute

For SI units,

$$CS = \frac{\pi DN}{1000} \text{ m/s} \tag{31}$$

where D is in millimeters and N is in revolutions per second.

The tool advancing rate is $F = f \times N$ ipm (mm/s). The machining time (T_1) required to turn a workpiece of length L in. (mm) is calculated from

$$T_1 = \frac{L}{F} \text{ min (s)} \tag{32}$$

The machining time (T_2) required to face a workpiece of diameter D is given by

$$T_2 = \frac{D/2}{F} \text{ min (s)} \tag{33}$$

The rate of metal removal (MRR) (Q) is given by

$$Q = 12fdCS \text{ in.}^3/\text{min} \tag{34}$$

Table 5 Typical Cutting Speeds ft/min (m/s)

Material	High-Speed Steel		Carbide	
	Finish ^a	Rough ^b	Finish ^a	Rough ^b
Free cutting steels, 1112, 1315	250–350 (1.3–1.8)	80–160 (0.4–0.8)	600–750 (3.0–3.8)	350–500 (1.8–2.5)
Carbon steels, 1010, 1025	225–300 (1.1–1.5)	80–130 (0.4–0.6)	550–700 (2.8–3.5)	300–450 (1.5–2.3)
Medium steels, 1030, 1050	200–300 (1.0–1.5)	70–120 (0.4–0.6)	450–600 (2.3–3.0)	250–400 (1.3–2.0)
Nickel steels, 2330	200–300 (1.0–1.5)	70–110 (0.4–0.6)	425–550 (2.1–2.8)	225–350 (1.1–1.8)
Chromium nickel, 3120, 5140	150–200 (0.8–1.0)	60–80 (0.3–0.4)	325–425 (1.7–2.1)	175–300 (0.9–1.5)
Soft gray cast iron	120–150 (0.6–0.8)	80–100 (0.4–0.5)	350–450 (1.8–2.3)	200–300 (1.0–1.5)
Brass, normal	275–350 (1.4–1.8)	150–225 (0.8–1.1)	600–700 (3.0–3.5)	400–600 (2.0–3.0)
Aluminum	225–350 (1.1–1.8)	100–150 (0.5–0.8)	450–700 (2.3–3.5)	200–350 (1.0–1.8)
Plastics	300–500 (1.5–2.5)	100–200 (0.5–1.0)	400–650 (2.0–3.3)	150–300 (0.8–1.5)

^aCut depth, 0.015–0.10 in. (0.38–2.54 mm); feed 0.005–0.015 ipr (0.13–0.38 mm/rev).

^bCut depth, 0.20–0.40 in. (5.0–10.0 mm); feed, 0.030–0.060 ipr (0.75–1.5 mm/rev).

$$\text{Power} = Q \text{ HP}_u \text{ HP} \quad (35)$$

$$\text{Power} = \text{torque } 2\pi N$$

$$= \frac{\text{torque} \times N}{63,025} \text{ HP} \quad (36)$$

where torque is in in.-lbf.

For SI units,

$$\text{Power} = \frac{\text{torque} \times N}{9549} \text{ kW} \quad (37)$$

where torque is in newton-meters and N in rev/min

$$\text{Torque} = F_c \times R$$

$$F_c = \frac{\text{torque}}{R} \quad (38)$$

where R is the radius of workpiece.

To convert to SI units,

$$\text{HP} \times 746 = ? \text{ watt (W)}$$

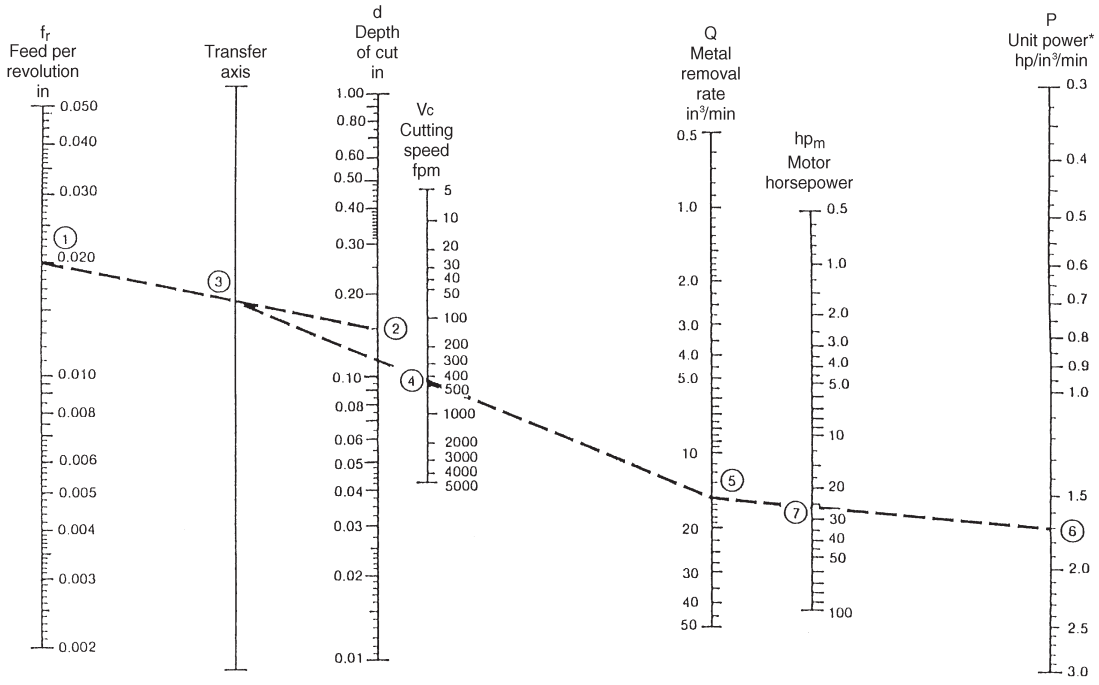
$$f(\text{lb}) \times 4.448 = ? \text{ newtons}$$

$$\text{Torque (in.-lb)} \times 0.11298 = ? \text{ newton-meter (N-m)}$$

$$\text{HP}/(\text{in.}^3/\text{min}) \times 2.73 = ? \text{ W}/(\text{mm}^3/\text{s})$$

$$\text{ft}/\text{min} \times 0.00508 = ? \text{ m/s}$$

$$\text{in.}^3 \times 16,390 = ? \text{ mm}^3$$



Example:
 $f_r = 0.020$ in
 $V_c = 425$ fpm
 $d = 0.150$ in
 $P = 1.70$ hp/in³/min

To Determine Motor Horsepower:

Connect Feed ① with Depth of cut ② to obtain point ③ on Transfer axis.
 Connect ③ with Cutting speed ④ to obtain point ⑤ on Metal removal rate scale.
 Connect point ⑤ with Unit power ⑥ to obtain 26 Motor horsepower at point ⑦.

Figure 11 Alignment chart for determining metal removal rate and motor horsepower in turning—English units.

Alignment charts were developed for determining metal removal rate and motor power in turning. Figures 11 and 12 show the method of using these charts either for English or metric units. The unit power (P) is the adjusted unit power with respect to turning conditions and machine efficiency.

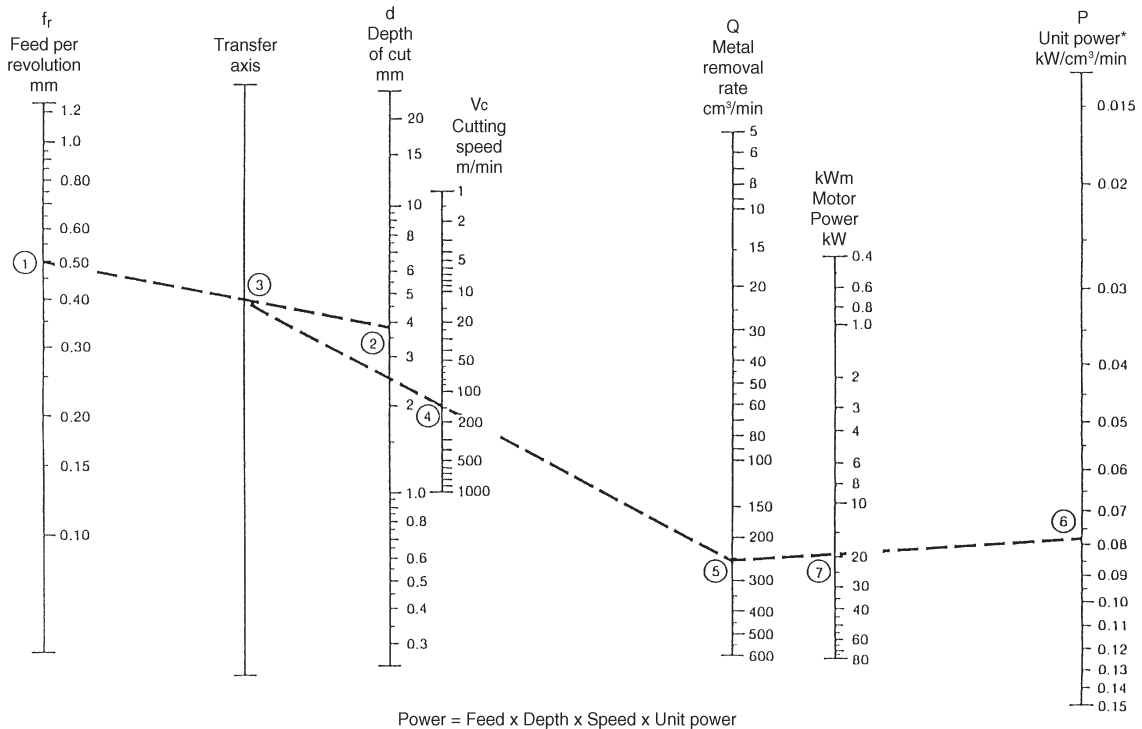
6.1 Lathe Size

The size of a lathe is specified in terms of the diameter of the work it will swing and the work-piece length it can accommodate. The main types of lathes are engine, turret, single-spindle automatic, automatic screw machine, multispindle automatic, multistation machines, boring, vertical, and tracer. The level of automation can range from semiautomatic to tape-controlled machining centers.

6.2 Break-Even Conditions

The selection of a specific machine for the production of a required quantity q must be done in a way to achieve minimum cost per unit produced. The incremental setup cost is given by DC_s , C_1 is the machining cost per unit on the first machine, and C_2 is the machining cost for the second machine, the break-even (BE) point will be calculated as follows:

$$BE = \Delta C \sqrt{C_1 - C_2}$$



Example:

$$\begin{aligned} f_r &= 0.5 \text{ mm} & d &= 3.8 \text{ mm} \\ V_c &= 130 \text{ m/min} & P &= 0.077 \text{ kW/cm}^3/\text{min} \end{aligned}$$

To Determine Motor Horsepower:

Connect Feed ① with Depth of cut ② to obtain point ③ on Transfer axis.

Connect ③ with Cutting speed ④ to obtain point ⑤ on Metal removal rate scale.

Connect point ⑤ with Unit power ⑥ to obtain 19.02 kW at Motor, point ⑦.

Figure 12 Alignment chart for determining metal removal rate and motor power in turning—metric units.

7 DRILLING MACHINES

Drills are used as the basic method of producing holes in a wide variety of materials. Figure 13 indicates the nomenclature of a standard twist drill and its comparison with a single-point tool. Knowledge of the thrust force and torque developed in the drilling process is important for design consideration. Figure 14 shows the forces developed during the drilling process. From the force diagram, the thrust force must be greater than $2P_y + P_y^1$ to include the friction on the sides and to be able to penetrate in the metal. The torque required is equal to P_2X . It is reported in the *Tool and Manufacturing Engineers Handbook*¹ that the following relations reasonably estimate the torque and thrust requirements of sharp twist drills of various sizes and designs.

$$\text{Torque: } M = Kf^{0.8}d^{1.8}A \text{ in.-lbf} \quad (39)$$

$$\text{Thrust: } T = 2Kf^{0.8}d^{0.8}B + kd^2E \text{ lb} \quad (40)$$

The thrust force has a large effect upon the required strength, rigidity, and accuracy, but the power required to feed the tool axially is very small.

$$\text{Cutting power: } \text{HP} = \frac{MN}{63,025} \quad (41)$$

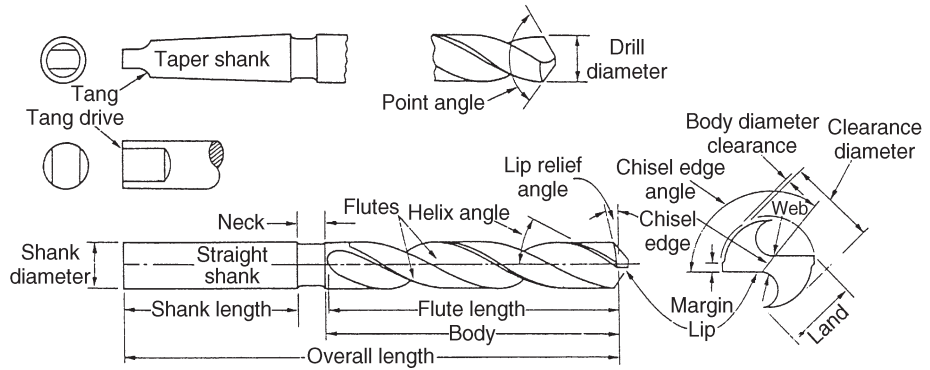


Figure 13 Drill geometry.

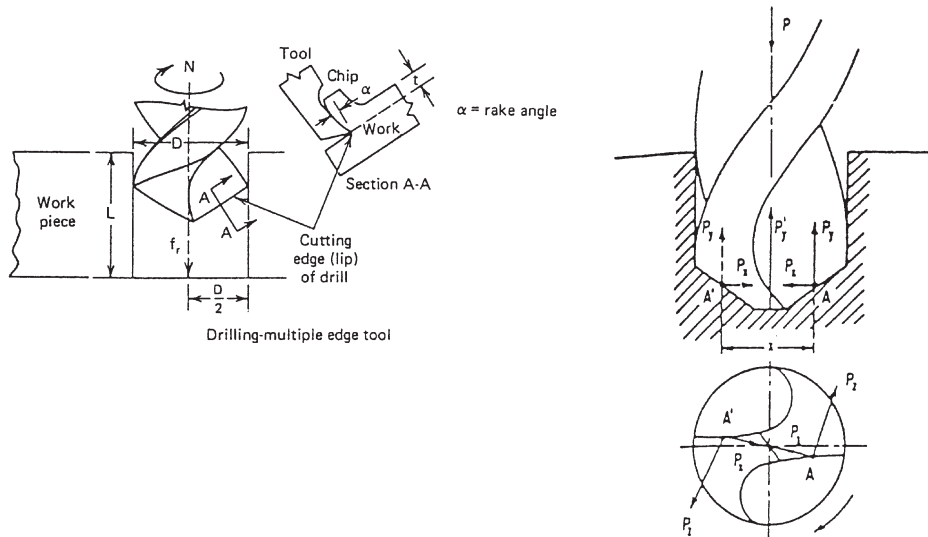


Figure 14 Thrust forces and torque in drilling operation.

where K = work-material constant
 f = drill feed, ipr
 d = drill diameter, in.
 A, B, E = design constants
 N = drill speed, rpm

Tables 6 and 7 give the constants used with the previous equations. Cutting speed at the surface is usually taken as 80% of turning speeds and is given by

$$CS = \frac{\pi d N}{12} \text{ fpm}$$

Force in cutting direction:

$$F_c = \frac{33,000 \text{ HP}}{CS} \text{ lb} \tag{42}$$

Table 6 Work-Material Constants for Calculating Torque and Thrust (National Twist Drill)

Work Material	K
Steel, 200 BHN	24,000
Steel, 300 BHN	31,000
Steel, 400 BHN	34,000
Most aluminum alloys	7,000
Most magnesium alloys	4,000
Most brasses	14,000
Leaded brass	7,000
Cast iron, 65 BHN	15,000
Free-machining mild steel, resulfurized	18,000
Austenitic stainless steel (type 316)	34,000

Table 7 Torque and Thrust Constants Based on Ratios c/d or w/d (National Twist Drill)

c/d	w/d	Torque Constant A	Thrust Constant B	Thrust Constant E
0.03	0.025	1.000	1.100	0.001
0.05	0.045	1.005	1.140	0.003
0.08	0.070	1.015	1.200	0.006
0.10	0.085	1.020	1.235	0.010
0.13	0.110	1.040	1.270	0.017
0.15	0.130	1.080	1.310	0.022
0.18	0.155	1.085	1.355	0.030
0.20	0.175	1.105	1.380	0.040
0.25	0.220	1.155	1.445	0.065
0.30	0.260	1.235	1.500	0.090
0.35	0.300	1.310	1.575	0.120
0.40	0.350	1.395	1.620	0.160

For SI units,

$$CS = \frac{\pi d_1 N}{60,000} \text{ m/s} \quad (43)$$

c = chisel-edge length, in.

d = drill diameter, in.

w = web thickness, in.

d_1 = drill diameter, in mm

Unit HP (hp/in.³/min) \times 2.73 = ? unit power (kW/cm³/s)

$$\text{kW} = \frac{MN}{9549} \quad (44)$$

M = torque Nm

For drills of regular proportion the ratio c/d is = 0.18 and $c = 1.15w$, approximately.

It is a common practice to feed drills at a rate that is proportional to the drill diameter in accordance with

$$f = \frac{d}{65} \quad (45)$$

Table 8 Recommended Feeds for Drills

Diameter		Feed	
in.	mm	ipr	mm/rev
Under 1/8	3.2	0.001–0.002	0.03–0.05
1/8–1/4	3.2–6.4	0.002–0.004	0.05–0.10
1/4–1/2	6.4–12.7	0.004–0.007	0.10–0.18
1/2–1	12.7–25.4	0.007–0.015	0.18–0.38
Over 1	25.4	0.015–0.025	0.38–0.64

For holes that are longer than $3d$, feed should be reduced. Also feeds and speeds should be adjusted due to differences in relative chip volume, material structure, cutting fluid effectiveness, depth of hole, and conditions of drill and machine. The advancing rate is

$$F = f \times N \text{ ipm} \quad (46)$$

The recommended feeds are given in Table 8.

The time T required to drill a hole of depth h is given by

$$T = \frac{h + 0.3d}{F} \text{ min} \quad (47)$$

The extra distance of $0.3d$ is approximately equal to the distance from the tip to the effective diameter of the tool. The rate of metal removal in case of blind holes is given by

$$Q = \left(\frac{\pi d^2}{4} \right) F \text{ in.}^3/\text{min} \quad (48)$$

When torque is unknown, the horsepower requirement can be calculated by

$$\text{HP}_c = Q \times C \times W \times (\text{HP}_u) \text{ hp}$$

C , W , HP_u are given in previous sections.

$$\text{Power} = \text{HP}_c \times 396,000 \text{ in.-lb/min} \quad (49)$$

$$\text{Torque} = \frac{\text{power}}{2\pi N} \text{ in.-lbf}$$

$$F_c = \frac{\text{torque}}{R} \text{ lb}$$

Along the cutting edge of the drill, the cutting speed is reduced toward the center as the diameter is reduced. The cutting speed is actually zero at the center. To avoid the region of very low speed and to reduce high thrust forces that might affect the alignment of the finished hole, a pilot hole is usually drilled before drilling holes of medium and large sizes. For the case of drilling with a pilot hole

$$\begin{aligned} Q &= \frac{\pi}{4} (d^2 - d_p^2)F \\ &= \frac{\pi}{4} (d + d_p)(d - d_p)F \text{ in.}^3/\text{min} \end{aligned} \quad (50)$$

Due to the elimination of the effects of the chisel-edge region, the equations for torque and thrust can be estimated as follows:

$$M_p = M \left[\frac{1 - (d_p/d)^2}{(1 + d_1/d)^{0.2}} \right] \quad (51)$$

$$T_p = T \left[\frac{1 - d_1/d}{(1 + d_1/d)^{0.2}} \right] \tag{52}$$

where d_p is the pilot hole diameter.

Alignment charts were developed for determining motor power in drilling. Figures 15 and 16 show the use of these charts either for English or metric units. The unit power (P^*) is the adjusted unit power with respect to drilling conditions and machine efficiency.

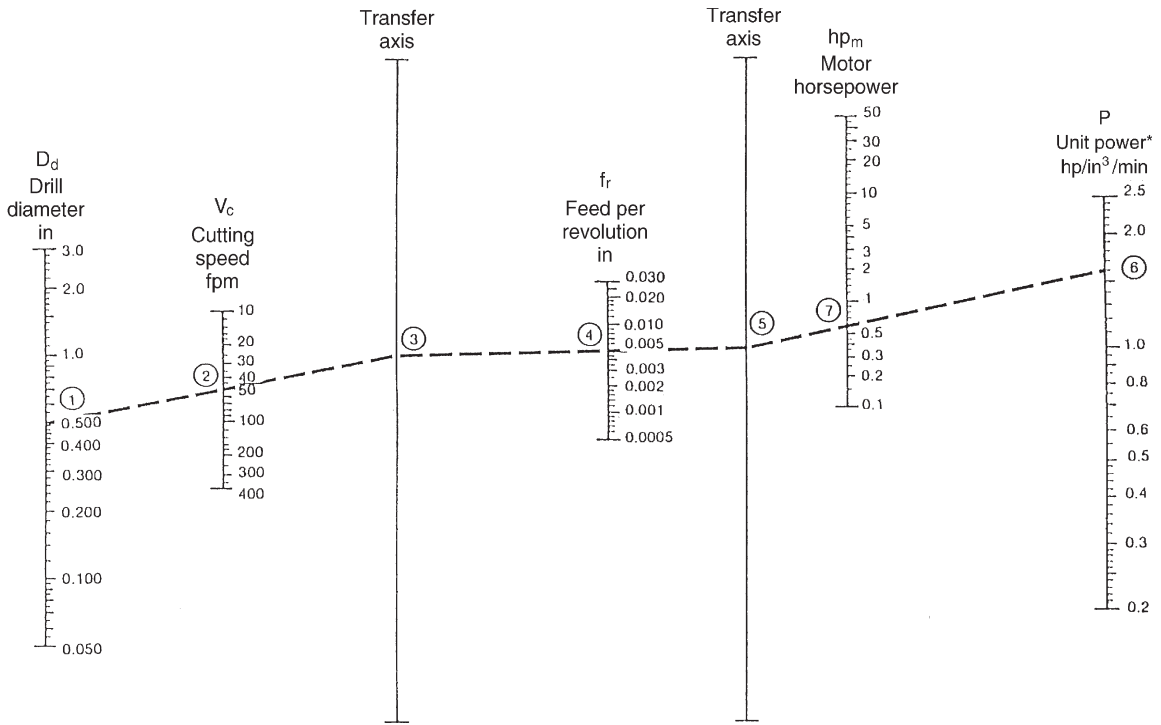
For English units,

$$HP_m = \frac{\pi D^2}{4} \times f \times N \times P^*$$

$$N = \frac{12V}{\pi D}$$

AS
$$HP_m = \frac{\pi D^2}{4} \times f \times \frac{12V}{\pi D} \times P^*$$

$$= 3D \times f \times V \times P^*$$



Example:

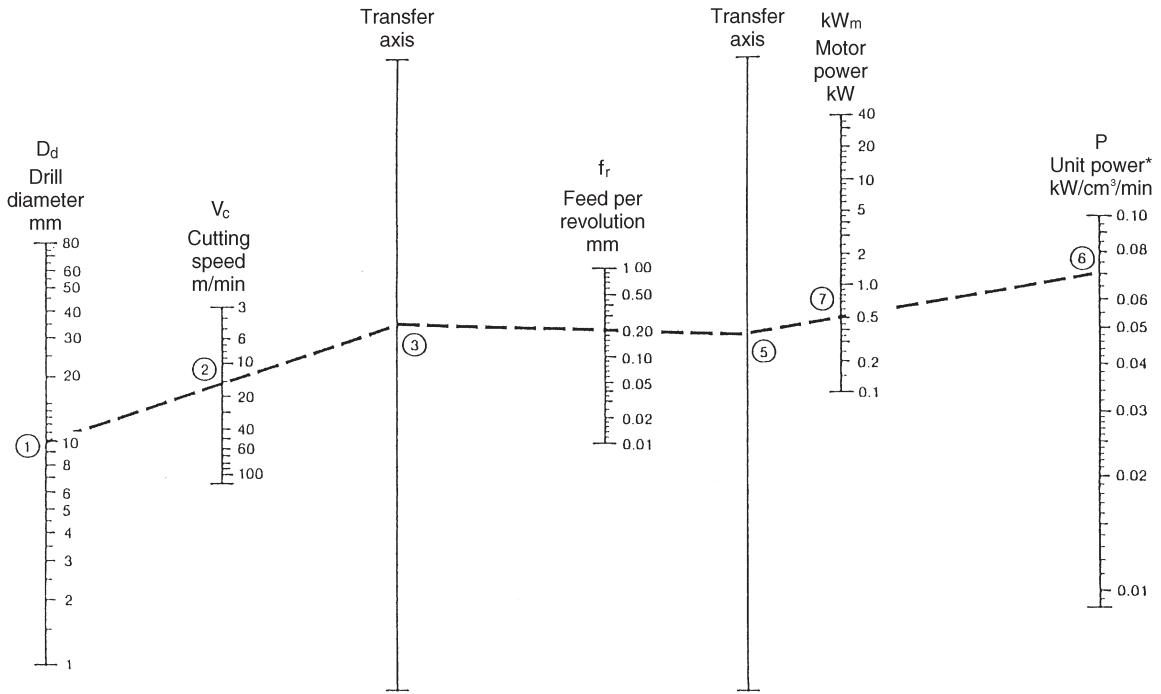
$D_d = 0.500$ in
 $V_c = 50$ fpm

$f_r = 0.005$ in
 $P = 1.6$ hp/in³/min

To Determine Motor Horsepower:

Connect Drill diameter ① with Cutting speed ② to obtain point ③ on Transfer axis. Connect ③ with Feed ④ to obtain point ⑤ on Transfer axis. Connect ⑤ with Unit power ⑥ to obtain 0.60 Motor horsepower, point ⑦.

Figure 15 Alignment chart for determining motor horsepower in drilling—English units.



$$\text{Motor Power} = 0.25 \times \text{Drill diameter} \times \text{Speed} \times \text{Feed} \times \text{Unit Power}$$

Example:

$D_d = 10 \text{ mm}$
 $V_c = 15 \text{ m/min}$

$f_r = 0.2 \text{ mm}$
 $P = 0.07 \text{ kW/cm}^3/\text{min}$

To Determine Motor Horsepower:

Connect Drill diameter ① with Cutting speed ② to obtain point ③ on Transfer axis. Connect ③ with Feed ④ to obtain point ⑤ on Transfer axis. Connect ⑤ with Unit power ⑥ to obtain 0.53 kW at motor power, point ⑦.

Figure 16 Alignment chart for determining motor power in drilling—metric units.

For metric units,

$$HP_m = \frac{\pi D^2}{4 \times 100} \times \frac{f}{10} \times N \times P \quad (P^* \text{ in kW/cm}^3/\text{min})$$

$$N = \frac{1000V}{\pi D}$$

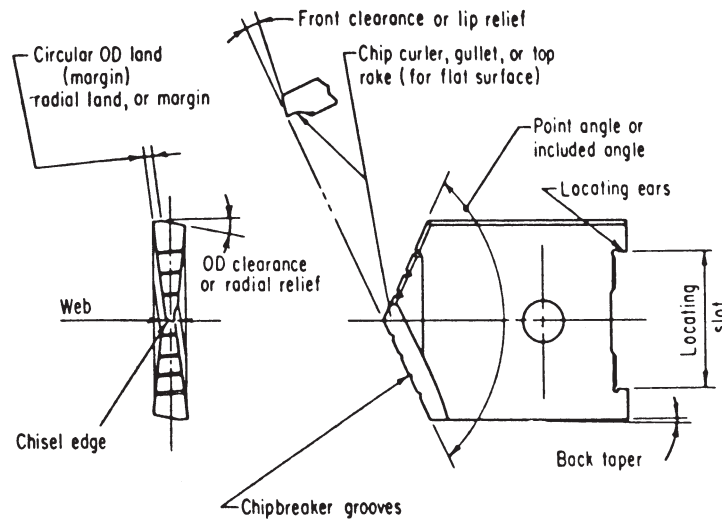
$$\begin{aligned} HP_m &= \frac{\pi D^2}{4 \times 100} \times \frac{f}{10} \times \frac{100V}{\pi D} \times P^* \\ &= 0.25D \times f \times V \times P^* \end{aligned}$$

7.1 Accuracy of Drills

The accuracy of holes drilled with a two-fluted twist drill is influenced by many factors, including the accuracy of the drill point; the size of the drill, the chisel edge, and the jigs used; the workpiece material; the cutting fluid used; the rigidity and accuracy of the machine used; and the cutting speed. Usually, when drilling most materials, the diameter of the drilled holes will be oversize. Table 9 provided the results of tests reported by The Metal Cutting Tool Institute for holes drilled in steel and cast iron.

Table 9 Oversize Diameters in Drilling

Drill Diameter (in.)	Amount Oversize (in.)		
	Average Max.	Mean	Average Min.
1/6	0.002	0.0015	0.001
1/8	0.0045	0.003	0.001
1/4	0.0065	0.004	0.0025
1/2	0.008	0.005	0.003
3/4	0.008	0.005	0.003
1	0.009	0.007	0.004

**Figure 17** Spade-drill blade elements.

Gun drills differ from conventional drills in that they are usually made with a single flute. A hole provides a passageway for pressurized coolant, which serves as a means of both keeping the cutting edge cool and flushing out the chips, especially in deep cuts.

Spade drills (Fig. 17) are made by inserting a spade-shaped blade into a shank. Some advantages of spade drills are (1) efficiency in making holes up to 15 in. in diameter; (2) low cost, since only the insert is replaced; (3) deep-hole drilling; and (4) easiness of chip breaking on removal.

Trepanning is a machining process for producing a circular hole, groove, disk, cylinder, or tube from solid stock. The process is accomplished by a tool containing one or more cutters, usually single-point, revolving around a center. The advantages of trepanning are (1) the central core left is solid material, not chips, which can be used in later work; and (2) the power required to produce a given hole diameter is highly reduced because only the annulus is actually cut.

Reaming, boring, counterboring, centering and countersinking, spotfacing, tapping, and chamfering processes can be done on drills. Microdrilling and submicrodrilling achieve holes in the range of 0.000025–0.20 in. in diameter.

Drilling machines are usually classified in the following manner:

1. Bench: plain or sensitive
2. Upright: single-spindle or turret
3. Radial
4. Gang
5. Multispindle
6. Deep hole: vertical or horizontal
7. Transfer

8 MILLING PROCESSES

The milling machines use a rotary multitooth cutter that can be designed to mill flat or irregularly shaped surfaces, cut gears, generate helical shapes, drill, bore, or do slotting work. Milling machines are classified broadly as vertical or horizontal. Figure 18 shows some of the operations that are done on both types.

Feed in milling (F) is specified in inches per minute, but it is determined from the amount each tooth can remove or feed per tooth (f_t). The feed in./min is calculated from

$$F = f_t \times n \times N \text{ in./min} \quad (53)$$

where n = number of teeth in cutter

$$N = \text{rpm}$$

Table 10 gives the recommended f_t for carbides and HSS tools. The cutting speed is calculated as follows:

$$CS = \frac{\pi DN}{12} \text{ fpm}$$

where D is the tool diameter, in.

Table 11 gives the recommended cutting speeds while using HSS and carbide-tipped tools. The relationship between cutter rotation and feed direction is shown in Fig. 19, while Fig. 20 shows the approach allowances required for both slot milling and face milling. In climb milling or down milling, the chips are cut to maximum thickness at initial engagement and decrease to zero thickness at the end of engagement. In conventional or up milling, the reverse occurs. Because of the initial impact, climb milling requires rigid machines with backlash eliminators.

The MRR is $Q = F \times w \times d$, where w is the width of cut and d is the depth of cut. The horsepower required for milling is given by

$$HP_c = HP_u \times Q$$

Machine horsepower is determined by

$$HP_m = \frac{HP_c}{\text{Eff.}} + HP_i \quad (54)$$

where HP_i is the idle horsepower.

Alignment charts were developed for determining MRR and motor power in face milling. Figures 21 and 22 show the method of using these charts either for English or metric units.

The time required for milling is equal to distance required to be traveled by the cutter to complete the cut (L_1) divided by the feed rate F . L_1 is equal to the length of cut (L) plus cutter approach A and the overtravel (OT). The machining time T is calculated from

$$T = \frac{L + A + OT}{F} \text{ min} \quad (55)$$

and OT depends on the specific milling operation.

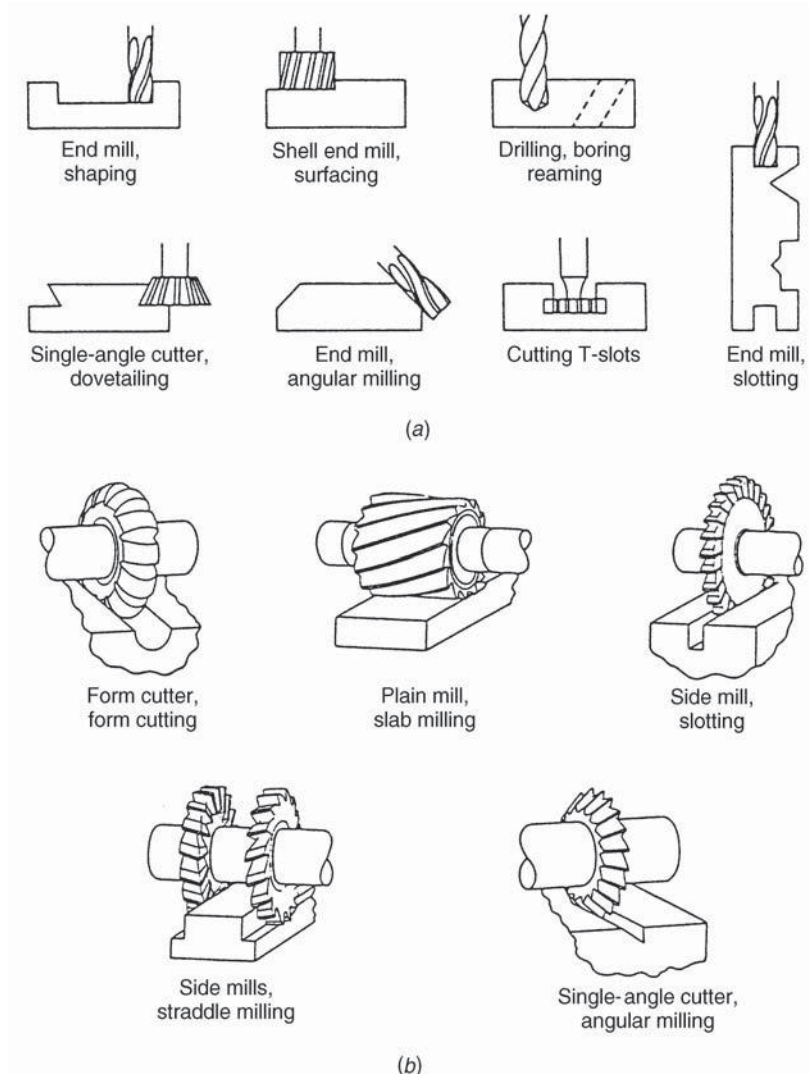


Figure 18 Applications of (a) vertical and (b) horizontal milling machines.

Table 10 Recommended Feed per Tooth for Milling Steel with Carbide and HSS Cutters

Type of Milling	Feed per Tooth	
	Carbides	HSS
Face	0.008–0.015	0.010
Side or straddle	0.008–0.012	0.006
Slab	0.008–0.012	0.008
Slotting	0.006–0.010	0.006
Slitting saw	0.003–0.006	0.003

Table 11 Table of Cutting Speeds (sfpm)–Milling

Work Material	HSS Tools		Carbide-Tipped Tools	
	Rough Mill	Finish Mill	Rough Mill	Finish Mill
Cast iron	50–60	80–110	180–200	350–400
Semisteel	40–50	65–90	140–160	250–300
Malleable iron	80–100	110–130	250–300	400–500
Cast steel	45–60	70–90	150–180	200–250
Copper	100–150	150–200	600	1000
Brass	200–300	200–300	600–1000	600–1000
Bronze	100–150	150–180	600	1000
Aluminum	400	700	800	1000
Magnesium	600–800	1000–1500	1000–1500	1000–5000
SAE steels				
1020 (coarse feed)	60–80	60–80	300	300
1020 (fine feed)	100–120	100–120	450	450
1035	75–90	90–120	250	250
X-1315	175–200	175–200	400–500	400–500
1050	60–80	100	200	200
2315	90–110	90–110	300	300
3150	50–60	70–90	200	200
4150	40–50	70–90	200	200
4340	40–50	60–70	200	200
Stainless steel	60–80	100–120	240–300	240–300
Titanium		30–70		200–350

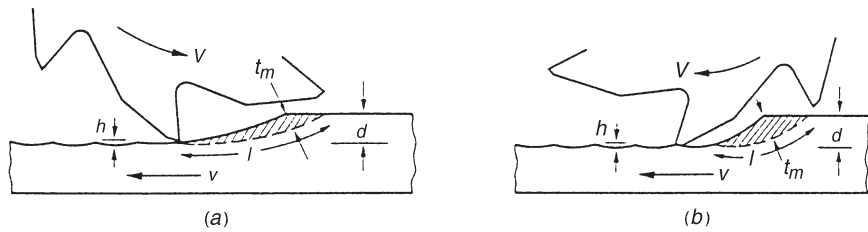


Figure 19 Cutting action in up-and-down milling.

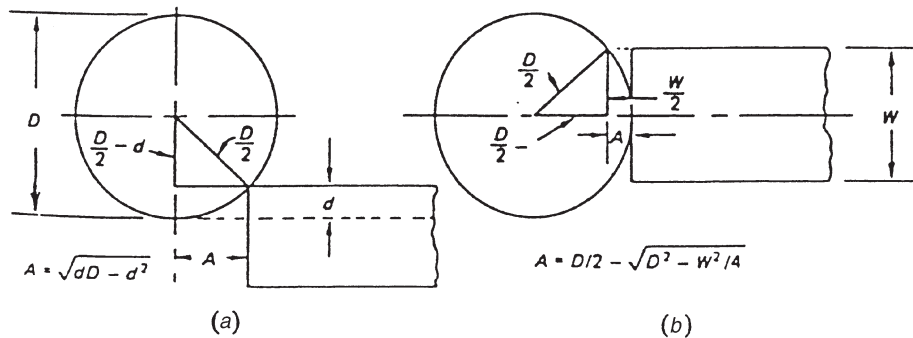
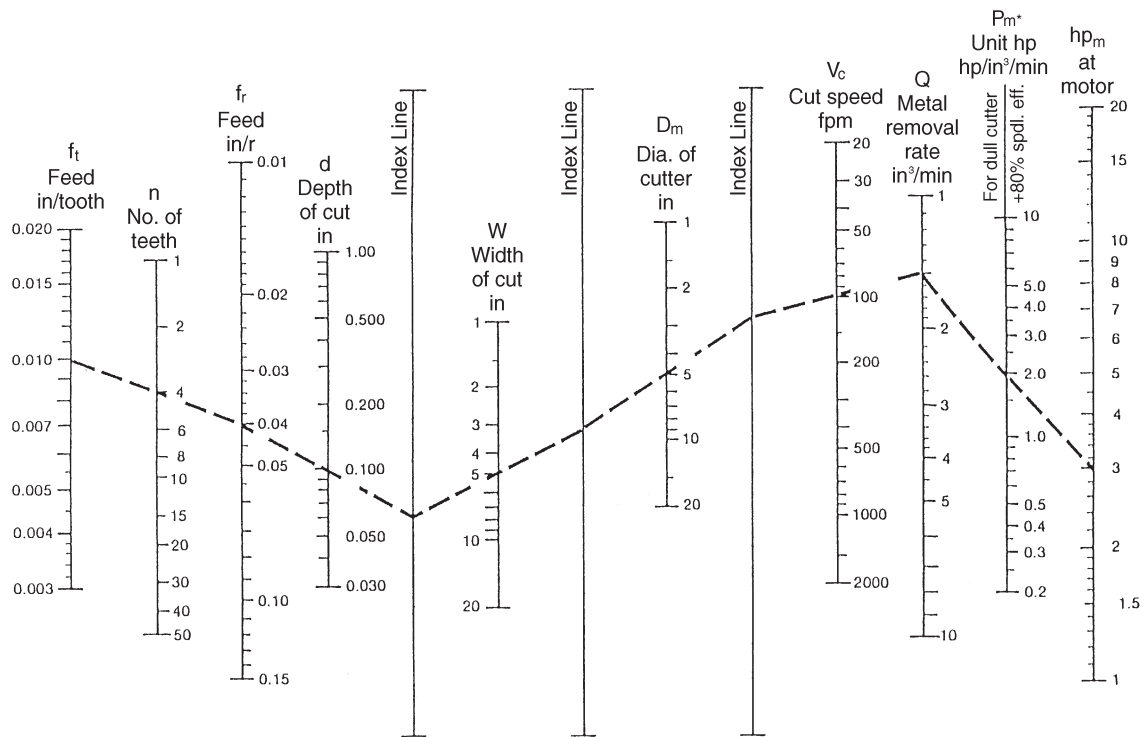


Figure 20 Allowance for approach in (a) plain or slot milling and (b) face milling.



Example:

$f_t = 0.010$ in/tooth
 $n = 4$ teeth
 $f_r = 0.04$ in/r

$d = 0.100$ in
 $w = 5$ in
 $D_m = 5$ in

$V_c = 100$ fpm
 $Q = 1.53$ in³/min
 $P = 2.0$ hp/in³/min
 $hp_m = 3.0$ hp

$$hp_m = \frac{P_m \times 3.82 \times f_t \times n \times d \times w \times V_c}{D_m}$$

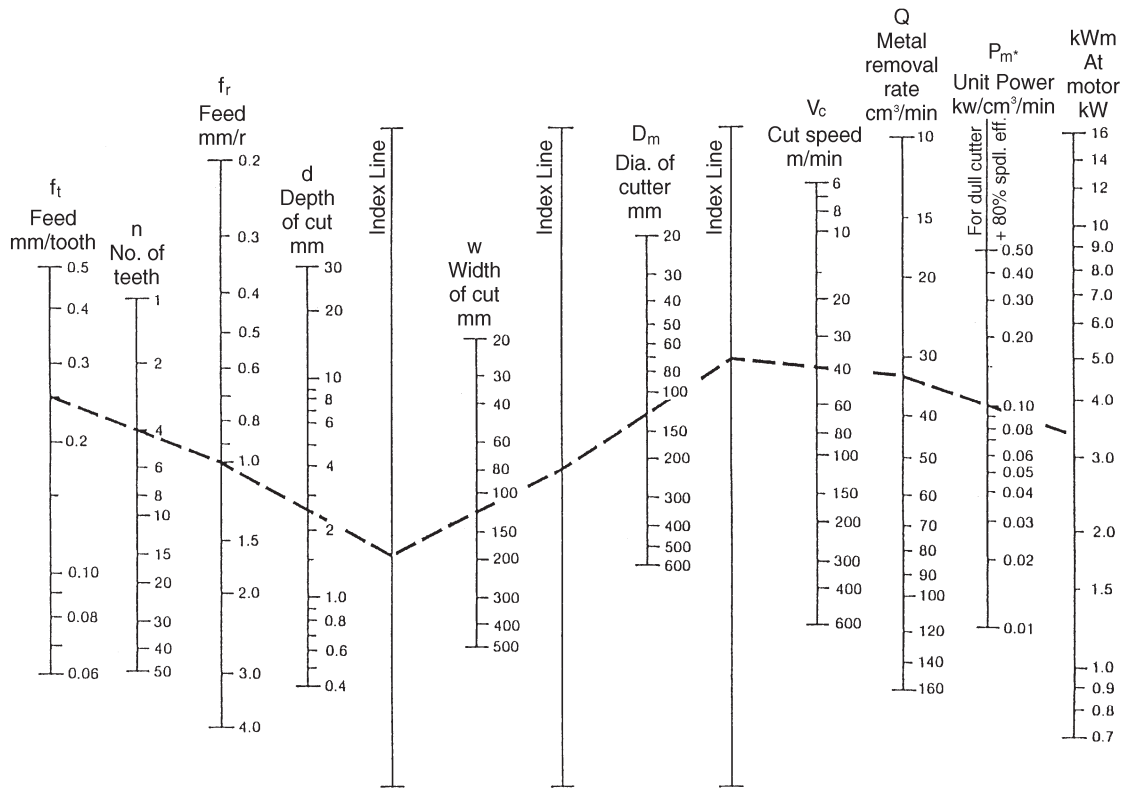
Figure 21 Alignment chart for determining metal removal rate and motor horsepower in face milling—English units.

The milling machines are designed according to the longitudinal table travel. Milling machines are built in different types, including

1. Column-and-knee: vertical, horizontal, universal, and ram
2. Bed type, multispindle
3. Planer
4. Special, turret, profilers, and duplicators
5. Numerically controlled

9 GEAR MANUFACTURING

Gears are made by various methods, such as machining, rolling, extrusion, blanking, powder metallurgy, casting, or forging. Machining still is the unsurpassed method of producing gears of all types and sizes with high accuracy. Roll forming can be used only on ductile materials; however, it has been highly developed and widely adopted in recent years. Casting, powder



Example:
 $f_t = 0.25$ mm/tooth $d = 2.5$ mm $V_c = 40$ m/min $hp_m = \frac{P_m \times 318.3 \times f_t \times n \times d \times w \times V_c}{1000 D_m}$
 $n = 4$ teeth $w = 125$ mm $Q = 31.8$ cm³/min
 $f_r = 1.0$ mm/r $D_m = 125$ mm $P = 0.1$ kW/cm³/min
 $kW_m = 3.18$ kW

Figure 22 Alignment chart for determining metal removal rate and motor power in face milling—metric units.

metallurgy, extruding, rolling, grinding, molding, and stamping techniques are used commercially in gear production.

9.1 Machining Methods

There are three basic methods for machining gears: form cutting, template machining, and the generating process.

Form cutting uses the principle illustrated in Fig. 23. The equipment and cutters required are relatively simple, and standard machines, usually milling, are often used. Theoretically, there should be different-shaped cutters for each size of gear for a given pitch, as there is a slight change in the curvature of the involute. However, one cutter can be used for several gears having different numbers of teeth without much sacrifice in their operating action. The eight standard involute cutters are listed in Table 12. On the milling machine, the index or dividing head is used to rotate the gear blank through a certain number of degrees after each cut. The rule to use is: turns of index handle = $40/N$, where N is the number of teeth. Form cutting is usually slow.

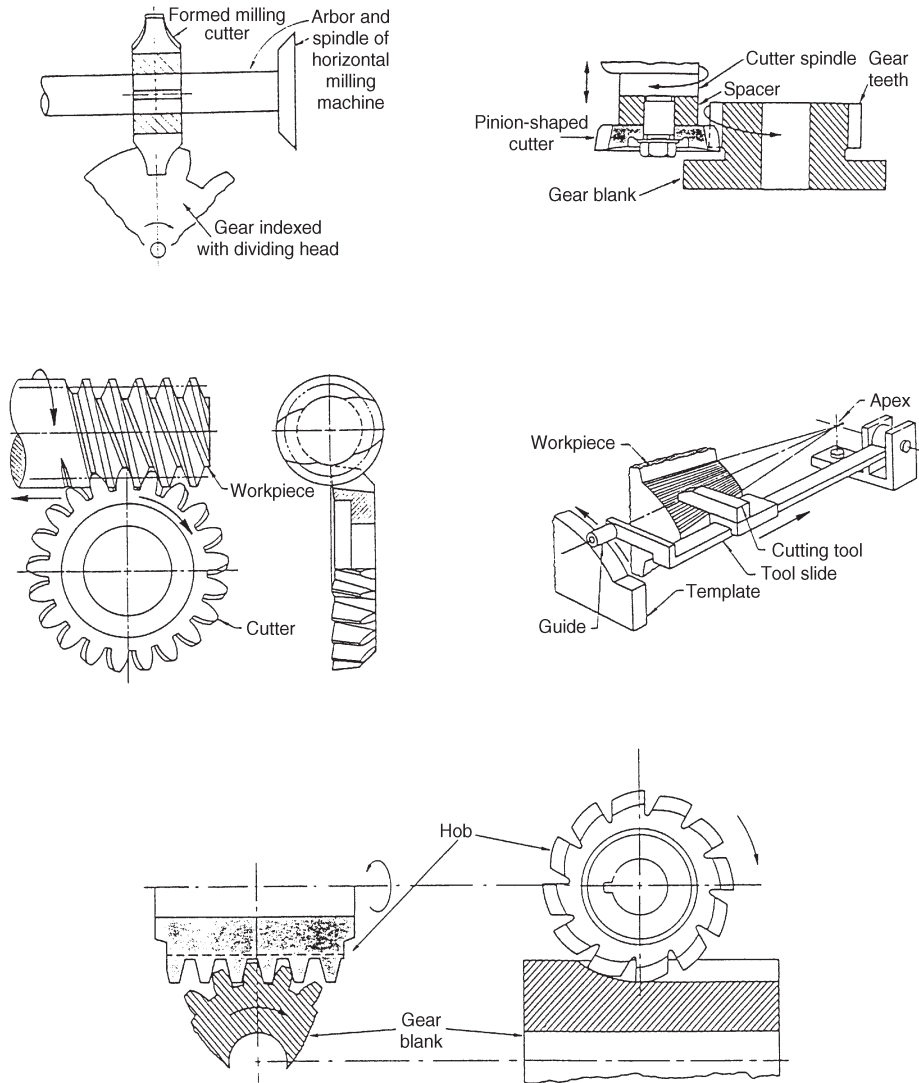


Figure 23 Machining gears.

Template machining utilizes a simple, single-point cutting tool that is guided by a template. However, the equipment is specialized, and the method is seldom used except for making large-bevel gears.

The *generating process* is used to produce most high-quality gears. This process is based on the principle that any two involute gears, or any gear and a rack, of the same diametral pitch will mesh together. Applying this principle, one of the gears (or the rack) is made into a cutter by proper sharpening and is used to cut into a mating gear blank and thus generate teeth on the blank. Gear shapers (pinion or rack), gear-hobbing machines, and bevel-gear generating machines are good examples of the gear generating machines.

Table 12 Standard Gear Cutters

Cutter Number	Gear Tooth Range
1	135 teeth to rack
2	55–34
3	35–54
4	26–34
5	21–25
6	17–20
7	14–16
8	12–13

9.2 Gear Finishing

To operate efficiently and have satisfactory life, gears must have accurate tooth profile and smooth and hard faces. Gears are usually produced from relatively soft blanks and are subsequently heat treated to obtain greater hardness, if it is required. Such heat treatment usually results in some slight distortion and surface roughness. *Grinding and lapping* are used to obtain very accurate teeth on hardened gears. Gear shaving and burnishing methods are used in gear finishing. Burnishing is limited to unhardened gears.

10 THREAD CUTTING AND FORMING

Three basic methods are used for the manufacturing of threads: *cutting, rolling, and casting*. Die casting and molding of plastics are good examples of casting. The largest number of threads are made by rolling, even though it is restricted to standardized and simple parts, and ductile materials. Large numbers of threads are cut by the following methods:

1. Turning
2. Dies: manual or automatic (external)
3. Milling
4. Grinding (external)
5. Threading machines (external)
6. Taps (internal)

10.1 Internal Threads

In most cases, the hole that must be made before an internal thread is tapped is produced by drilling. The hole size determines the depth of the thread, the forces required for tapping, and the tap life. In most applications, a drill size is selected that will result in a thread having about 75% of full thread depth. This practice makes tapping much easier, increases the tap's life, and only slightly reduces the resulting strength. Table 13 gives the drill sizes used to produce 75% thread depth for several sizes of UNC threads. The feed of a tap depends on the lead of the screw and is equal to $1/\text{lead ipr}$.

Cutting speeds depend on many factors such as:

1. Material hardness
2. Depth of cut

Table 13 Recommended Tap-Drill Sizes for Standard Screw-Thread Pitches (American National Coarse-Thread Series)

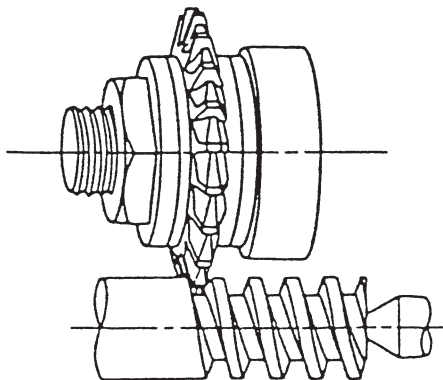
Number or Diameter	Threads per Inch	Outside Diameter of Screw	Tap Drill Sizes	Decimal Equivalent of Drill
6	32	0.138	36	0.1065
8	32	0.164	29	0.1360
10	24	0.190	25	0.1495
12	24	0.216	16	0.1770
1/4	20	0.250	7	0.2010
3/8	16	0.375	5/16	0.3125
1/2	13	0.500	27/64	0.4219
3/4	10	0.750	21/32	0.6562
1	8	1.000	7/8	0.875

3. Thread profile
4. Tooth depth
5. Hole depth
6. Fineness of pitch
7. Cutting fluid

Cutting speeds can range from lead 3 ft/min (1 m/min) for high-strength steels to 150 ft/min (45 m/min) for aluminum alloys. Long-lead screws with different configurations can be cut successfully on milling machines, as in Fig. 24. The feed per tooth is given by the following equation:

$$f_t = \frac{\pi dS}{nN} \quad (56)$$

where d = diameter of thread
 n = number of teeth in cutter
 N = rpm of cutter
 S = rpm of work

**Figure 24** Single-thread milling cutter.

10.2 Thread Rolling

In thread rolling, the metal on the cylindrical blank is cold forged under considerable pressure by either rotating cylindrical dies or reciprocating flat dies. The advantages of thread rolling include improved strength, smooth surface finish, less material used (~19%), and high production rate. The limitations are that blank tolerance must be close, it is economical only for large quantities, it is limited to external threads, and it is applicable only for ductile materials, less than Rockwell C37.

11 BROACHING

Broaching is unique in that it is the only one of the basic machining processes in which the feed of the cutting edges is built into the tool. The machined surface is always the inverse of the profile of the broach. The process is usually completed in a single, linear stroke. A broach is composed of a series of single-point cutting edges projecting from a rigid bar, with successive edges protruding farther from the axis of the bar. Figure 25 illustrates the parts and nomenclature of the broach. Most broaching machines are driven hydraulically and are of the pull or push type.

The maximum force an internal pull broach can withstand without damage is given by

$$P = \frac{A_y F_y}{s} \text{ lb} \tag{57}$$

where A_y = minimum tool selection, in.²
 F_y = tensile yield strength of tool steel, psi
 s = safety factor

The maximum push force is determined by the minimum tool diameter (D_y), the length of the broach (L), and the minimum compressive yield strength (F_y). The ratio L/D_y should be

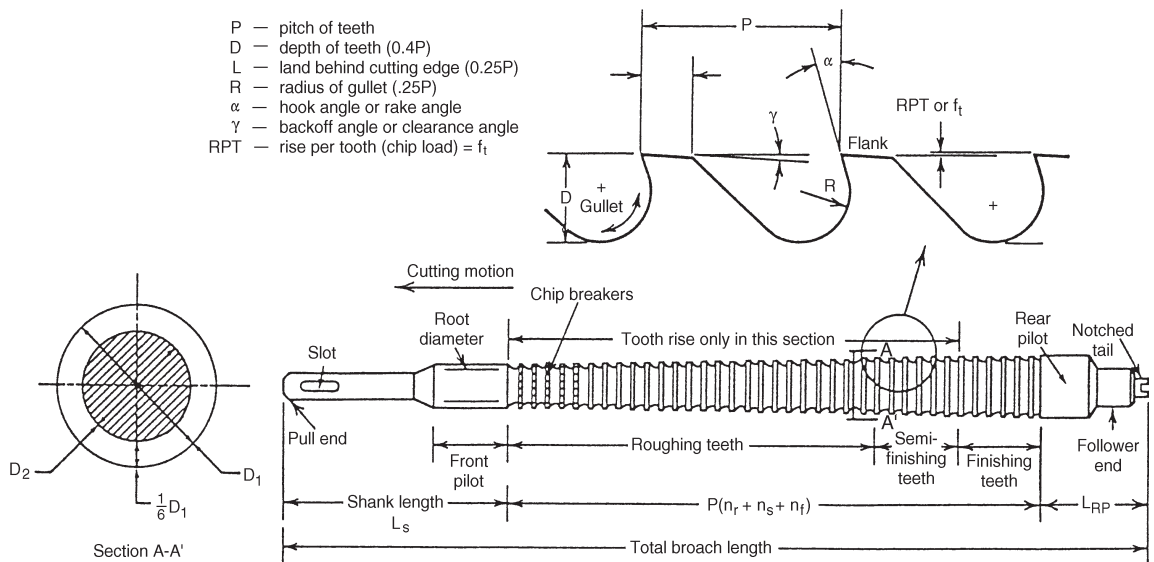


Figure 25 Standard broach part and nomenclature.

less than 25 so that the tool will not bend under load. The maximum allowable pushing force is given by

$$P = \frac{A_y F_y}{s} \text{ lb} \quad (58)$$

where F_y is minimum compressive yield strength.

If L/D_y ratio is greater than 25 (long broach), the *Tool and Manufacturing Engineers Handbook* gives the following formula:

$$P = \frac{5.6 \times 10^7 D_r^4}{sL^2} \text{ lb} \quad (59)$$

where D_r and L are given in inches.

Alignment charts were developed for determining MRR and motor power in surface broaching. Figures 26 and 27 show the application of these charts for either English or metric units.

Broaching speeds are relatively low, seldom exceeding 50 fpm, but, because a surface is usually completed in one stroke, the productivity is high.

12 SHAPING, PLANING, AND SLOTTING

The shaping and planing operations generate surfaces with a single-point tool by a combination of a reciprocating motion along one axis and a feed motion normal to that axis (Fig. 28). Slots and limited inclined surfaces can also be produced. In shaping, the tool is mounted on a reciprocating ram, and the table is fed at each stroke of the ram. Planers handle large, heavy workpieces. In planing, the workpiece reciprocates and the feed increment is provided by moving the tool at each reciprocation. To reduce the lost time on the return stroke, they are provided with a quick-return mechanism. For mechanically driven shapers, the ratio of cutting time to return stroke averages 3:2, and for hydraulic shapers the ratio is 2:1. The average cutting speed may be determined by the following formula:

$$CS = \frac{LN}{12C} \text{ fpm} \quad (60)$$

where N = strokes per minute

L = stroke length, in.

C = cutting time ratio, cutting time divided by total time

For mechanically driven shapers, the cutting speed reduces to

$$CS = \frac{LN}{7.2} \text{ fpm} \quad (61)$$

or

$$CS = \frac{L_1 N}{600} \text{ m/min} \quad (62)$$

where L_1 is the stroke length in millimeters. For hydraulically driven shapers,

$$CS = \frac{LN}{8} \text{ fpm} \quad (63)$$

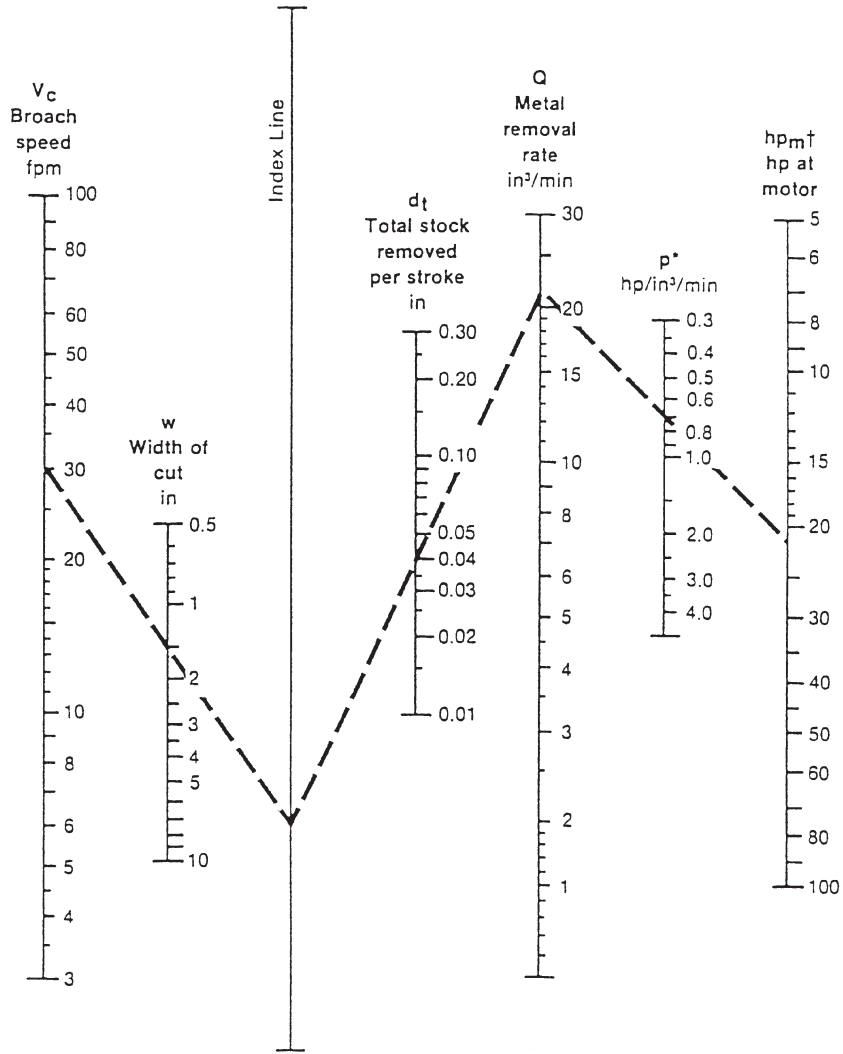
or

$$CS = \frac{L_1 N}{666.7} \text{ m/min} \quad (64)$$

The time T required to machine a workpiece of width W (in.) is calculated by

$$T = \frac{W}{N \times f} \text{ min} \quad (65)$$

where f is the feed, in. per stroke.



Example:

Material: Cast iron — HSS tools

Chipload 0.005 in/tooth

$V_c = 30$ fpm

$w = 1.5$ in

$d_t = 0.040$ in

$Q = 22$ in³/min

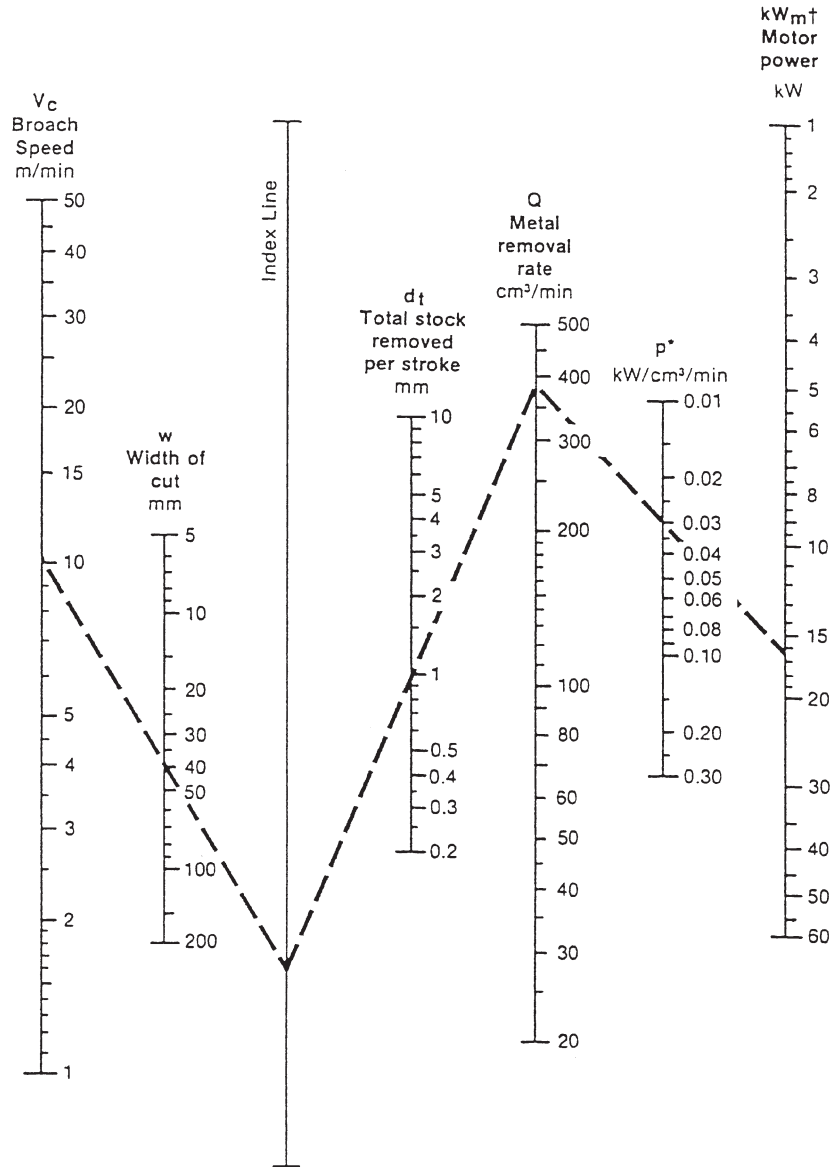
$P = 0.7$ hp/in³/min

$hp_m = 22$ hp

$$Q = 12 V_c \times w \times d_t \text{ in}^3/\text{min}$$

$$hp_m = \frac{Q \times P}{E} = \frac{Q \times P}{0.7}$$

Figure 26 Alignment chart for determining metal removal rate and motor horsepower in surface broaching with high-speed steel broaching tools—English units.



Example:

Material: Cast iron — HSS tools

Chipload 0.13 mm/tooth

$V_c = 10$ m/min

$w = 38$ mm

$d_t = 1$ mm

$Q = 380$ cm³/min

$P = 0.03$ kW/cm³/min

$P_m = 16.3$ kW

$$Q = V_c \times w \times d_t \text{ cm}^3/\text{min}$$

$$P_m = \frac{Q \times P}{E} = \frac{Q \times P}{0.7}$$

Figure 27 Alignment chart for determining metal removal rate and motor power in surface broaching with high-speed steel broaching tools—metric units.

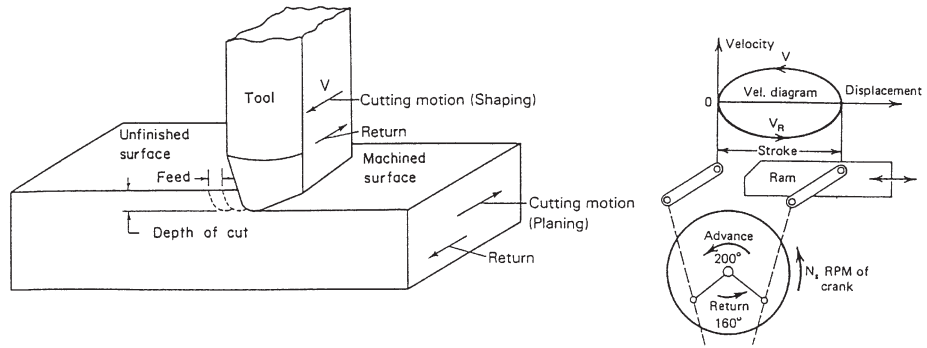


Figure 28 Basic relationships of tool motion, feed, and depth of cut in shaping and planing.

The number of strokes (S) required to complete a job is then

$$S = \frac{W}{f} \tag{66}$$

The power required can be approximated by

$$HP_c = K df(CS) \tag{67}$$

where d = depth of cut, in.

CS = cutting speed, fpm

K = cutting constant, for medium cast iron, 3; free-cutting steel, 6; and bronze, 1.5
or

$$HP_c = 12f \times d \times CS \times HP_\mu$$

$$F_c = \frac{33,000 HP_c}{CS}$$

13 SAWING, SHEARING, AND CUTTING OFF

Saws are among the most common of machine tools, even though the surfaces they produce often require further finishing operations. Saws have two general areas of applications: contouring and cutting off. There are three basic types of saws: hacksaw, circular, and band saw.

The *reciprocating power hacksaw* machines can be classified as either positive or uniform-pressure feeds. Most of the new machines are equipped with a quick-return action to reduce idle time.

The machining time required to cut a workpiece of width W in. is calculated as follows:

$$T = \frac{W}{fN} \text{ min} \tag{68}$$

where F = feed, in./stroke

N = number of strokes per min

Circular saws are made of three types: metal saws, steel friction disks, and abrasive disks. Solid metal saws are limited in size, not exceeding 16 in. in diameter. Large circular saws have either replaceable inserted teeth or segmented-type blades. The machining time required to cut a workpiece of width W in. is calculated as follows:

$$T = \frac{W}{f_1 n N} \text{ min} \tag{69}$$

where f_t = feed per tooth
 n = number of teeth
 N = rpm

Steel friction disks operate at high peripheral speeds ranging from 18,000 to 25,000 fpm (90–125 m/s). The heat of friction quickly softens a path through the part. The disk, which is sometimes provided with teeth or notches, pulls and ejects the softened metal. About 0.5 min are required to cut through a 24-in. I-beam.

Abrasive disks are mainly aluminum oxide grains or silicon carbide grains bonded together. They will cut ferrous or nonferrous metals. The finish and accuracy is better than steel friction blades, but they are limited in size compared to steel friction blades.

Band saw blades are of the continuous type. Band sawing can be used for cutting and contouring. Band-sawing machines operate with speeds that range from 50 to 1500 fpm. The time required to cut a workpiece of width W in. can be calculated as follows:

$$T = \frac{W}{12f_t n V} \text{ min} \quad (70)$$

where f_t = feed, in. per tooth
 n = number of teeth per in.
 V = cutting speed, fpm

Cutting can also be achieved by band-friction cutting blades with a surface speed up to 15,000 fpm. Other band tools include band filing, diamond bands, abrasive bands, spiral bands, and special-purpose bands.

14 MACHINING PLASTICS

Most plastics are readily formed, but some machining may be required. Plastic's properties vary widely. The general characteristics that affect their machinability are discussed below.

First, all plastics are poor heat conductors. Consequently, little of the heat that results from chip formation will be conducted away through the material or carried away in the chips. As a result, cutting tools run very hot and may fail more rapidly than when cutting metal. Carbide tools frequently are more economical to use than HSS tools if cuts are of moderately long duration or if high-speed cutting is to be done.

Second, because considerable heat and high temperatures do develop at the point of cutting, thermoplastics tend to soften, swell, and bind or clog the cutting tool. Thermosetting plastics give less trouble in this regard.

Third, cutting tools should be kept very sharp at all times. Drilling is best done by means of straight-flute drills or by "dubbing" the cutting edge of a regular twist drill to produce a zero rake angle. Rotary files and burrs, saws, and milling cutters should be run at high speeds to improve cooling, but with feed carefully adjusted to avoid jamming the gullets. In some cases, coolants can be used advantageously if they do not discolor the plastic or cause gumming. Water, soluble oil and water, and weak solutions of sodium silicate in water are used. In turning and milling plastics, diamond tools provide the best accuracy, surface finish, and uniformity of finish. Surface speeds of 500–600 fpm with feeds of 0.002–0.005 in. are typical.

Fourth, filled and laminated plastics usually are quite abrasive and may produce a fine dust that may be a health hazard.

15 GRINDING, ABRASIVE MACHINING, AND FINISHING

Abrasive machining is the basic process in which chips are removed by very small edges of abrasive particles, usually synthetic. In many cases, the abrasive particles are bonded into

wheels of different shapes and sizes. When wheels are used mainly to produce accurate dimensions and smooth surfaces, the process is called *grinding*. When the primary objective is rapid metal removal to obtain a desired shape or approximate dimensions, it is termed *abrasive machining*. When fine abrasive particles are used to produce very smooth surfaces and to improve the metallurgical structure of the surface, the process is called *finishing*.

15.1 Abrasives

Aluminum oxide (Al_2O_3), usually synthetic, performs best on carbon and alloy steels, annealed malleable iron, hard bronze, and similar metals. Al_2O_3 wheels are not used in grinding very hard materials, such as tungsten carbide, because the grains will get dull prior to fracture. Common trade names for aluminum oxide abrasives are *Alundum* and *Aloxite*.

Silicon carbide (SiC), usually synthetic, crystals are very hard, being about 9.5 on the Moh scale, where diamond hardness is 10. SiC crystals are brittle, which limits their use. Silicon carbide wheels are recommended for materials of low tensile strength, such as cast iron, brass, stone, rubber, leather, and cemented carbides.

Cubic boron nitride (CBN) is the second hardest natural or man-made substance. It is good for grinding hard and tough-hardened tool-and-die steels.

Diamonds may be classified as natural or synthetic. Commercial diamonds are now manufactured in high, medium, and low impact strength.

Grain Size

To have uniform cutting action, abrasive grains are graded into various sizes, indicated by the numbers 4–600. The number indicates the number of openings per linear inch in a standard screen through which most of the particles of a particular size would pass. Grain sizes 4–24 are termed coarse; 30–60, medium; and 70–600, fine. Fine grains produce smoother surfaces than coarse ones but cannot remove as much metal.

Bonding materials have the following effects on the grinding process: (1) they determine the strength of the wheel and its maximum speed; (2) they determine whether the wheel is rigid or flexible; and (3) they determine the force available to pry the particles loose. If only a small force is needed to release the grains, the wheel is said to be soft. Hard wheels are recommended for soft materials and soft wheels for hard materials. The bonding materials used are vitrified, silicate, rubber, resinoid, shellac, and oxychloride.

Structure or Grain Spacing

Structure relates to the spacing of the abrasive grain. Soft, ductile materials require a wide spacing to accommodate the relatively large chips. A fine finish requires a wheel with a close spacing. Figure 29 shows the standard system of grinding wheels as adopted by the American National Standards Institute.

Speeds

Wheel speed depends on the wheel type, bonding material, and operating conditions. Wheel speeds range between 4500 and 18,000 sfpm (22.86 and 27.9 m/s); and 5500 sfpm (27.9 m/s) is generally recommended as best for all disk-grinding operations. Work speeds depend on type of material, grinding operation, and machine rigidity. Work speeds range between 15 and 200 fpm.

Feeds

Cross feed depends on the width of grinding wheel. For rough grinding, the range is one-half to three-quarters of the width of the wheel. Finer feed is required for finishing, and it ranges between one-tenth and one-third of the width of the wheel. A cross feed between 0.125 and 0.250 in. is generally recommended.

Depth of Cut

Rough-grinding conditions will dictate the maximum depth of cut. In the finishing operation, the depth of cut is usually small, 0.0002–0.001 in. (0.005–0.025 mm). Good surface finish and close tolerance can be achieved by “sparking out” or letting the wheel run over the workpiece without increasing the depth of cut till sparks die out. The *grinding ratio* (*G* ratio) refers to the ratio of the cubic inches of stock removed to the cubic inches of grinding wheel worn away. *G* ratio is important in calculating grinding and abrasive machining cost, which may be calculated by the following formula:

$$C = \frac{C_a}{G} + \frac{L}{tq} \quad (71)$$

where *C* = specific cost of removing a in.³ of material

*C*_a = cost of abrasive, \$/in.³

G = grinding ratio

L = labor and overhead charge, \$/h

q = machining rate, in.³/h

t = fraction of time the wheel is in contact with workpiece

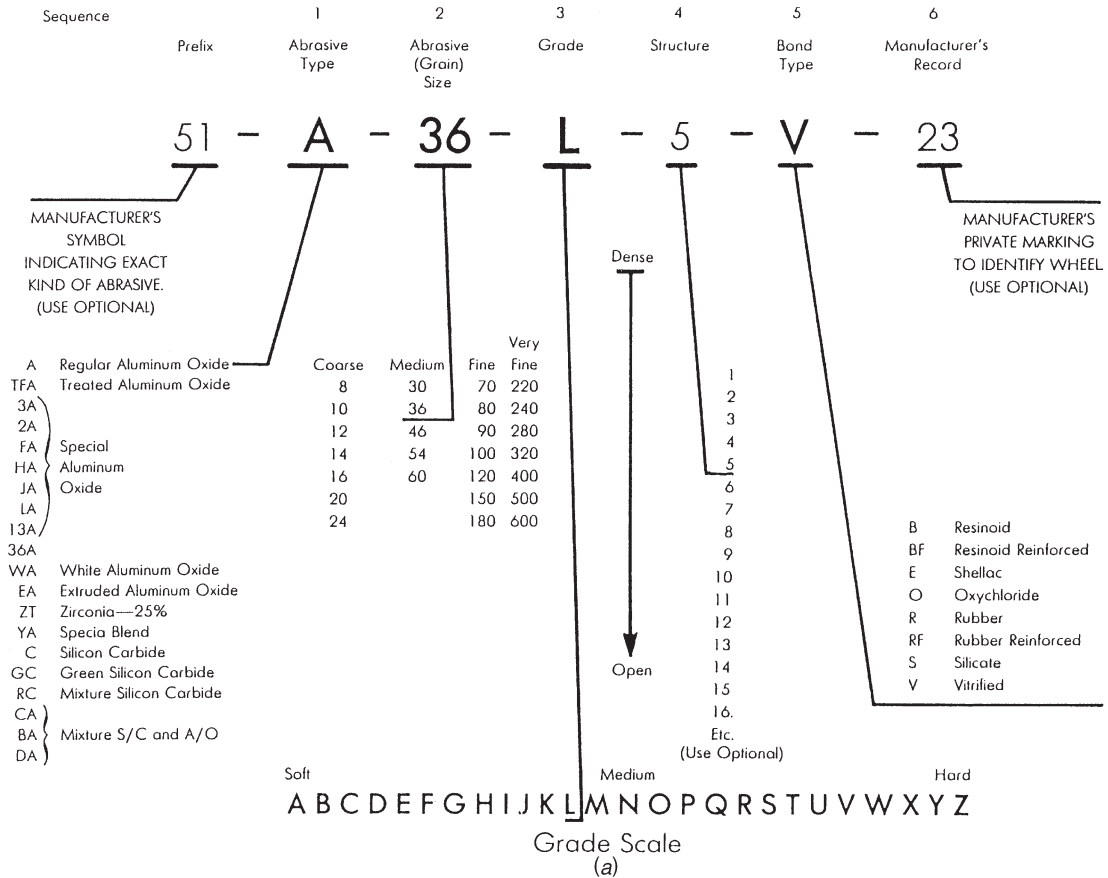
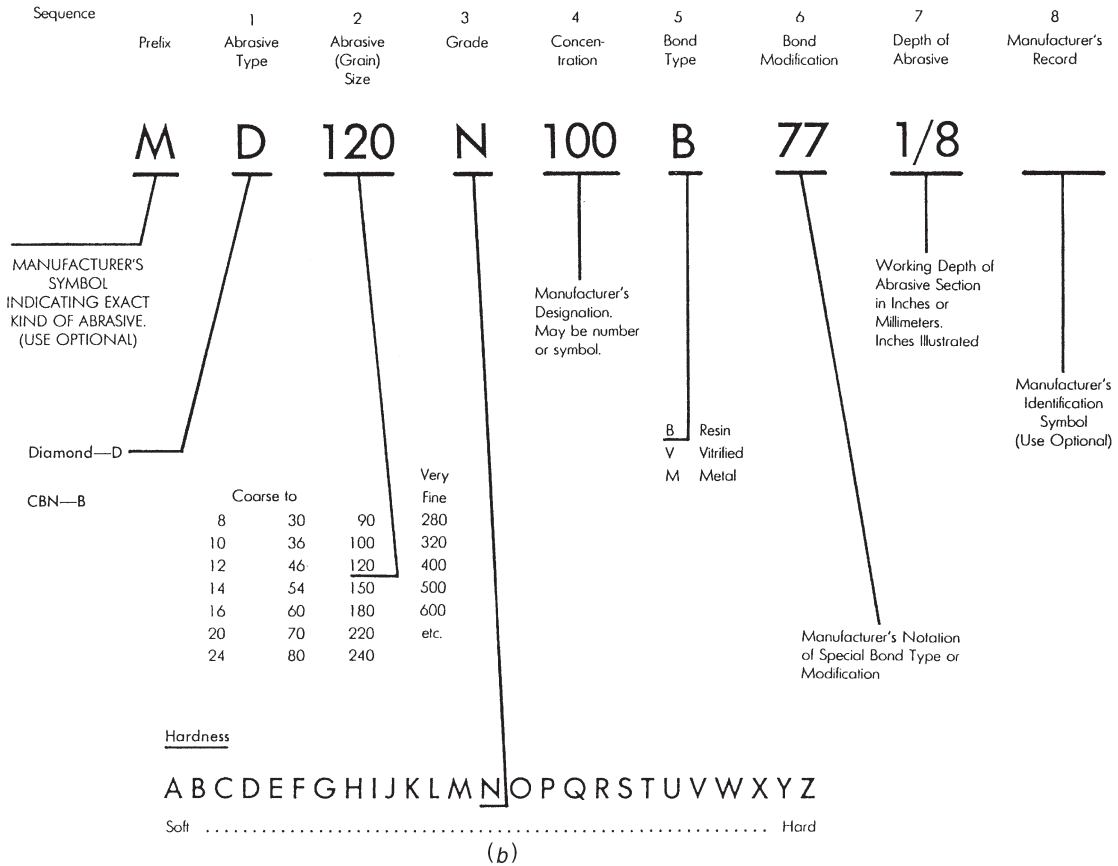


Figure 29 Standard systems for grinding wheels. (a) aluminum oxide, silicon carbide; and (b) diamond, CBN.



(b) Figure 29 (continued)

Power Requirement

$$\text{Power} = (u)(\text{MRR}) = F_c \times R \times 2\pi N$$

$$\text{MRR} = \text{material removal rate} = d \times w \times v$$

- where d = depth of cut
- w = width of cut
- v = work speed
- u = specific energy for surface grinding. Table 14 gives the approximate specific energy requirement for certain metals
- R = radius of wheel
- N = rev/unit time

15.2 Temperature

Temperature rise affects the surface properties and causes residual stresses on the workpiece. It is related to process variables by the following relation:

$$\text{Temperature rise} \propto D^{1/4} d^{3/4} \left(\frac{V}{v}\right)^{1/2} \tag{72}$$

Table 14 Approximate Specific Energy Required for Surface Grinding

Workpiece Material	Hardness	HP(in. ³ /min)	W/(mm ³ /s)
Aluminum	150 HB	3–10	8–27
Steel	(110–220) HB	6–24	16–66
Cast iron	(140–250) HB	5–22	14–60
Titanium alloy	300 HB	6–20	16–55
Tool steel	62–67 HRC	7–30	19–82

where D = wheel diameter

V = wheel speed

Grinding Fluids

Grinding fluids are water-based emulsions for general grinding and oils for thread and gear grinding. Advantages include:

1. Machining hard materials > RC50.
2. Fine surface finish, 10–80 $\mu\text{in.}$ (0.25–2 μm).
3. Accurate dimensions and close tolerances 1.0002 in. (1.005 mm) can be easily achieved.
4. Grinding pressure is light.

Machines

Grinding and abrasive machines include:

1. Surface grinders, reciprocating or rotating table
2. Cylindrical grinders, work between centers, centerless, crankshaft, thread and gear form work, and internal and other special applications
3. Jig grinders
4. Tool and cutter grinders
5. Snagging, foundry rough work
6. Cutting off and profiling
7. Abrasive grinding, belt, disk and loose grit
8. Mass media, barrel tumbling, and vibratory

Ultrasonic Machining

In ultrasonic machining, material is removed from the workpiece by microchipping or erosion through high-velocity bombardment by abrasive particles, in the form of a slurry, through the action of an ultrasonic transducer. It is used for machining hard and brittle materials and can produce very small and accurate holes 0.015 in. (0.4 mm).

Surface Finishing

Finishing processes produce an extra-fine surface finish; in addition, tool marks are removed and very close tolerances are achieved. Some of these processes follow.

Honing is a low-velocity abrading process. It uses fine abrasive stones to remove very small amounts of metals usually left from previous grinding processes. The amount of metal removed

is usually less than 0.005 in. (0.13 mm). Because of low cutting speeds, heat and pressure are minimized, resulting in excellent sizing and metallurgical control.

Lapping is an abrasive surface-finishing process wherein fine abrasive particles are charged in some sort of a vehicle, such as grease, oil, or water, and are embedded into a soft material, called a *lap*. Metal laps must be softer than the work and are usually made of close-grained gray cast iron. Other materials, such as steel, copper, and wood, are used where cast iron is not suitable. As the charged lap is rubbed against a surface, small amounts of material are removed from the harder surface. The amount of material removed is usually less than 0.001 in. (0.03 mm).

Superfinishing is a surface-improving process that removes undesirable fragmentation, leaving a base of solid crystalline metal. It uses fine abrasive stones, like honing, but differs in the type of motion. Very rapid, short strokes, very light pressure, and low-viscosity lubricant-coolant are used in superfinishing. It is essentially a finishing process and not a dimensional one and can be superimposed on other finishing operations.

Buffing

Buffing wheels are made from a variety of soft materials. The most widely used is muslin, but flannel, canvas, sisal, and heavy paper are used for special applications. Buffing is usually divided into two operations: cutting down and coloring. The first is used to smooth the surface and the second to produce a high luster. The abrasives used are extremely fine powders of aluminum oxide, tripoli (an amorphous silicon), crushed flint or quartz, silicon carbide, and red rouge (iron oxide). Buffing speeds range between 6000 and 12,000 fpm.

Electropolishing is the reverse of electroplating; that is, the work is the anode instead of the cathode and metal is removed rather than added. The electrolyte attacks projections on the workpiece surface at a higher rate, thus producing a smooth surface.

16 NONTRADITIONAL MACHINING

Nontraditional, or nonconventional, machining processes are material removal processes that have recently emerged or are new to the user. They have been grouped for discussion here according to their primary energy mode; that is, mechanical, electrical, thermal, or chemical, as shown in Table 15.

Nontraditional processes provide manufacturing engineers with additional choices or alternatives to be applied where conventional processes are not satisfactory, such as when

- Shapes and dimensions are complex or very small.
- Hardness of material is very high (>400 HB).
- Tolerances are tight and very fine surface finish is desired.
- Temperature rise and residual stresses must be avoided.
- Cost and production time must be reduced.

Figure 30 and Table 16 demonstrate the relationships among the conventional and the nontraditional machining processes with respect to surface roughness, dimensional tolerance, and metal-removal rate. The *Machinery Handbook*⁶ is an excellent reference for nontraditional machining processes, values, ranges, and limitations.

16.1 Abrasive Flow Machining

Abrasive flow machining (AFM) is the removal of material by a viscous, abrasive medium flowing, under pressure, through or across a workpiece. Figure 31 contains a schematic presentation of the AFM process. Generally, the puttylike medium is extruded through or over

Table 15 Current Commercially Available Nontraditional Material Removal Processes

Mechanical		Electrical		Thermal		Chemical	
AFM	Abrasive flow machining	ECD	Electrochemical deburring	EBM	Electron-beam machining	CHM	Chemical machining: chemical milling, chemical blanking
AJM	Abrasive jet machining	ECDG	Electrochemical discharge grinding	EDG	Electrical discharge grinding		
HDM	Hydrodynamic machining						
LSG	Low-stress grinding	ECG	Electrochemical grinding	EDM	Electrical discharge machining	ELP	Electropolish
RUM	Rotary ultrasonic machining	ECH	Electrochemical honing			PCM	Photochemical machining
		ECM	Electrochemical machining	EDS	Electrical discharge sawing	TCM	Thermochemical machining (or TEM, thermal energy method)
TAM	Thermally assisted machining	ECP	Electrochemical polishing	EDWC	Electrical discharge wire cutting		
		ECS	Electrochemical sharpening				
TFM	Total form machining	ECT	Electrochemical turning	LBM	Laser beam machining		
USM	Ultrasonic machining	ES	Electrostream&	LBT	Laser beam torch		
WJM	Water-jet machining	STEM TM	Shaped tube electrolytic machining	PBM	Plasma beam machining		

the workpiece with motion usually in both directions. Aluminum oxide, silicon carbide, boron carbide, or diamond abrasives are used. The movement of the abrasive matrix erodes away burrs and sharp corners and polishes the part.

16.2 Abrasive Jet Machining

Abrasive jet machining (AJM) is the removal of material through the action of a focused, high-velocity stream of fine grit or powder-loaded gas. The gas should be dry, clean, and under modest pressure. Figure 32 shows a schematic of the AJM process. The mixing chamber sometimes uses a vibrator to promote a uniform flow of grit. The hard nozzle is directed close to the workpiece at a slight angle.

16.3 Hydrodynamic Machining

Hydrodynamic machining (HDM) removes material by the stroking of high-velocity fluid against the workpiece. The jet of fluid is propelled at speeds up to Mach 3. Figure 33 shows a schematic of the HDM operation.

16.4 Low-Stress Grinding

Low-stress grinding (LSG) is an abrasive material-removal process that leaves a low-magnitude, generally compressive, residual stress on the surface of the workpiece. Figure 34 shows a

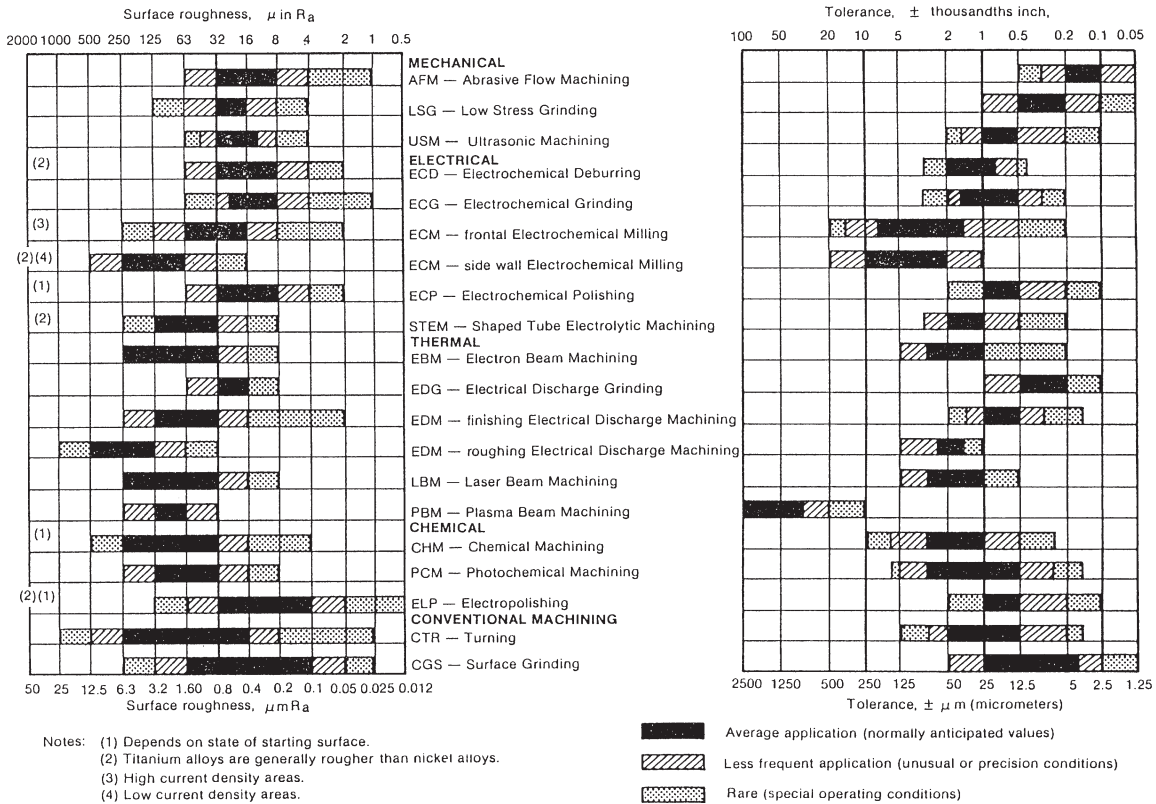


Figure 30 Typical surface roughness and tolerances produced by nontraditional machining.

schematic of the LSG process. The thermal effects from conventional grinding can produce high tensile stress in the workpiece surface. The process parameter guidelines can be applied to any of the grinding modes: surface, cylindrical, centerless, internal, and so on.

16.5 Thermally Assisted Machining

Thermally assisted machining (TAM) is the addition of significant amounts of heat to the workpiece immediately prior to single-point cutting so that the material is softened but the strength of the tool bit is unimpaired (Fig. 35). While resistive heating and induction heating offer possibilities, the plasma arc has a core temperature of 14,500°F (8000°C) and a surface temperature of 6500°F (3600°C). The torch can produce 2000°F (1100°C) in the workpiece in approximately one-quarter revolution of the workpiece between the point of application of the torch and the cutting tool.

16.6 Electromechanical Machining

Electromechanical machining (EMM) is a process in which the metal removal is affected in a conventional manner except that the workpiece is electrochemically polarized. When the applied voltage and the electrolytic solution are controlled, the surface of the workpiece can be changed to achieve the characteristics suitable for the machining operation.

Table 16 Material Removal Rates and Dimensional Tolerances

Process	Maximum Rate of Material Removal in. ³ /min (cm ³ /min)	Typical Power Consumption hp/in. ³ /min (kW/cm ³ /min)	Cutting Speed fpm (m/min)	Penetration Rate per Minute in. (mm)	Accuracy ±		Typical Machine Input hp (kW)
					Attainable in. (mm)	At Maximum Material Removal Rate in. (mm)	
Conventional turning	200	1	250	—	0.0002	0.005	30
Conventional grinding	3300	0.046	76	—	0.005	0.13	22
	50	10	10	—	0.0001	0.002	25
CHM	820	0.46	3	—	0.0025	0.05	20
	30	—	—	0.001	0.0005	0.003	—
PBM	490	—	—	0.025	0.013	0.075	—
	10	20	50	10	0.02	0.1	200
ECG	164	0.91	15	254	0.5	2.54	150
	2	2	0.25	—	0.0002	0.0025	4
ECM	33	0.019	0.08	—	0.005	0.063	3
	1	160	—	0.5	0.0005	0.006	200
EDM	16.4	7.28	—	12.7	0.013	0.15	150
	0.3	40	—	0.5	0.00015	0.002	15
USM	4.9	1.82	—	12.7	0.004	0.05	11
	0.05	200	—	0.02	0.0002	0.0015	15
EBM	0.82	9.10	—	0.50	0.005	0.040	11
	0.0005	10,000	200	6	0.0002	0.002	10
LBM	0.0082	455	60	150	0.005	0.050	7.5
	0.0003	60,000	—	4	0.0005	0.005	20
	0.0049	2,731	—	102	0.013	0.13	15

16.7 Total Form Machining

Total form machining (TFM) is a process in which an abrasive master abrades its full three-dimensional shape into the workpiece by the application of force while a full-circle, orbiting motion is applied to the workpiece via the worktable (Fig. 36). The cutting master is advanced into the work until the desired depth of cut is achieved. Uniformity of cutting is promoted by the fluid that continuously transports the abraded particles out of the working gap. Adjustment of the orbiting cam drive controls the precision of the overcut from the cutting master. Cutting action takes place simultaneously over the full surface of abrasive contact.

16.8 Ultrasonic Machining

Ultrasonic machining (USM) is the removal of material by the abrading action of a grit-loaded liquid slurry circulating between the workpiece and a tool vibrating perpendicular to the work face at a frequency above the audible range (Fig. 37). A high-frequency power source activates a stack of magnetostrictive material, which produces a low-amplitude vibration of the tool holder. This motion is transmitted under light pressure to the slurry, which abrades the workpiece into a conjugate image of the tool form. A constant flow of slurry (usually cooled) is necessary to carry away the chips from the work face. The process is sometimes called *ultrasonic abrasive machining* (UAM) or *impact machining*.

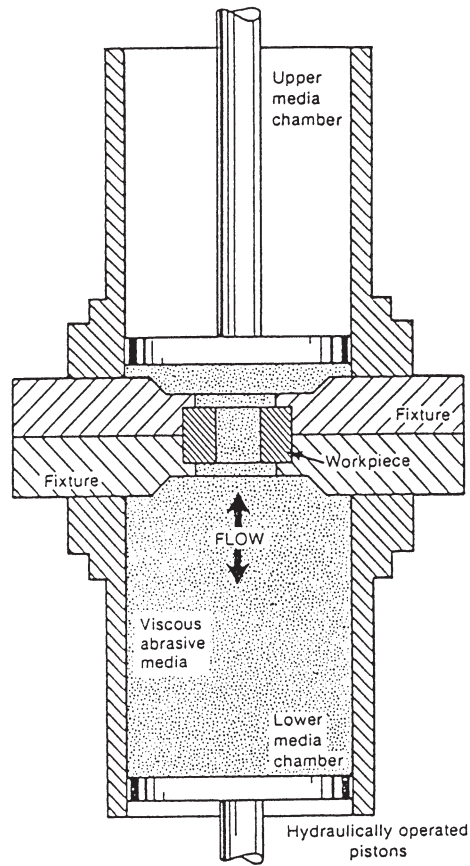


Figure 31 Abrasive flow machining.

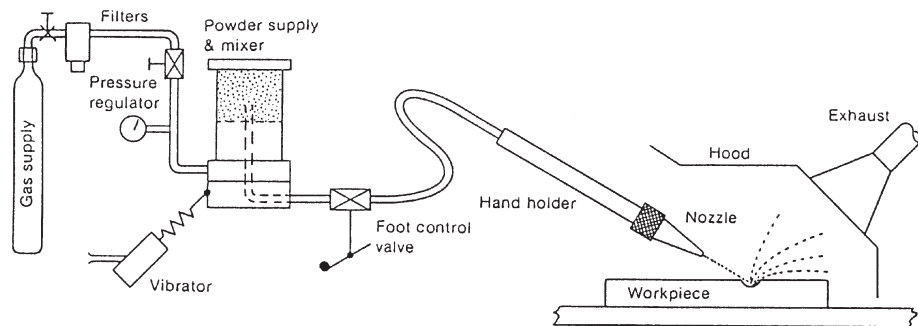


Figure 32 Abrasive jet machining.

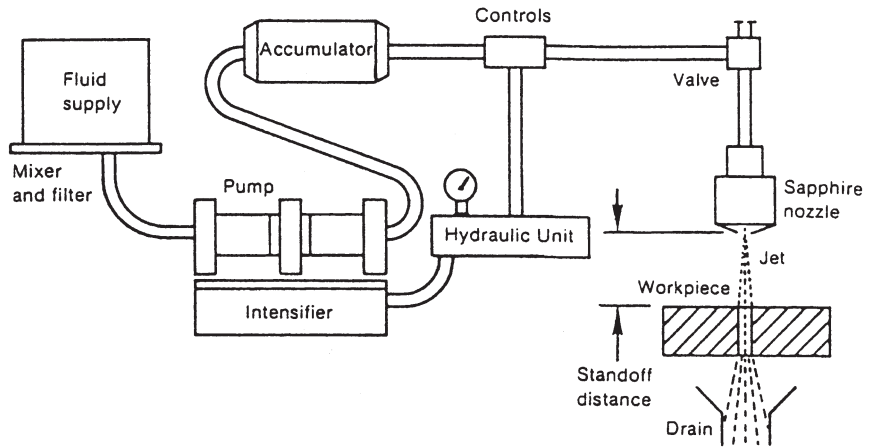


Figure 33 Hydrodynamic machining.

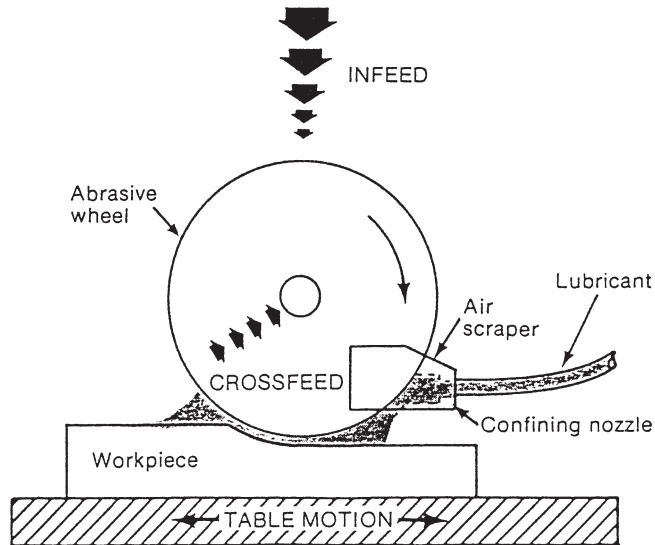


Figure 34 Low-stress grinding.

A prime variation of USM is the addition of ultrasonic vibration to a rotating tool—usually a diamond-plated drill. *Rotary ultrasonic machining* (RUM) substantially increases the drilling efficiency. A piezoelectric device built into the rotating head provides the needed vibration. Milling, drilling, turning, threading, and grinding-type operations are performed with RUM.

16.9 Water-Jet Machining

Water-jet machining (WJM) is low-pressure hydrodynamic machining. The pressure range for WJM is an order of magnitude below that used in HDM. There are two versions of WJM: one for

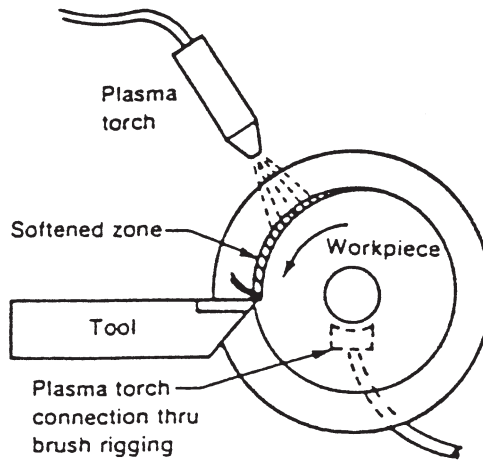


Figure 35 Thermally assisted machining.

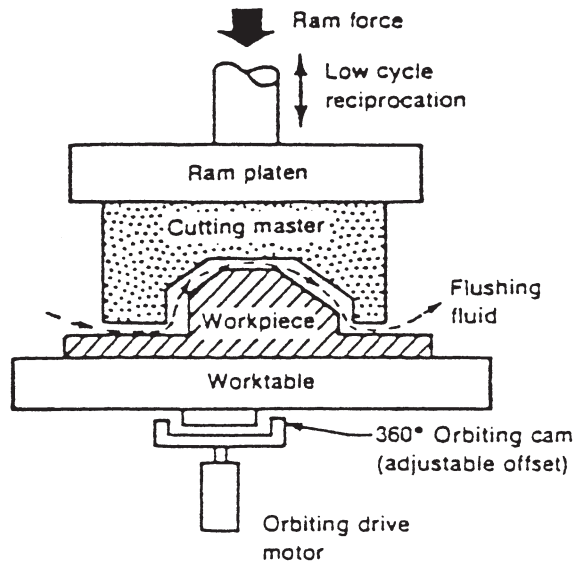


Figure 36 Total form machining.

mining, tunneling, and large-pipe cleaning that operates in the region from 250 to 1000 psi (1.7–6.9 MPa); and one for smaller parts and production shop situations that uses pressures below 250 psi (1.7 MPa).

The first version, or high-pressure range, is characterized by use of a pumped water supply with hoses and nozzles that generally are hand directed. In the second version, more production-oriented and controlled equipment, such as that shown in Fig. 38, is involved. In some instances, abrasives are added to the fluid flow to promote rapid cutting. Single or multiple-nozzle approaches to the workpiece depend on the size and number of parts per load. The principle is that WJM is high volume, not high pressure.

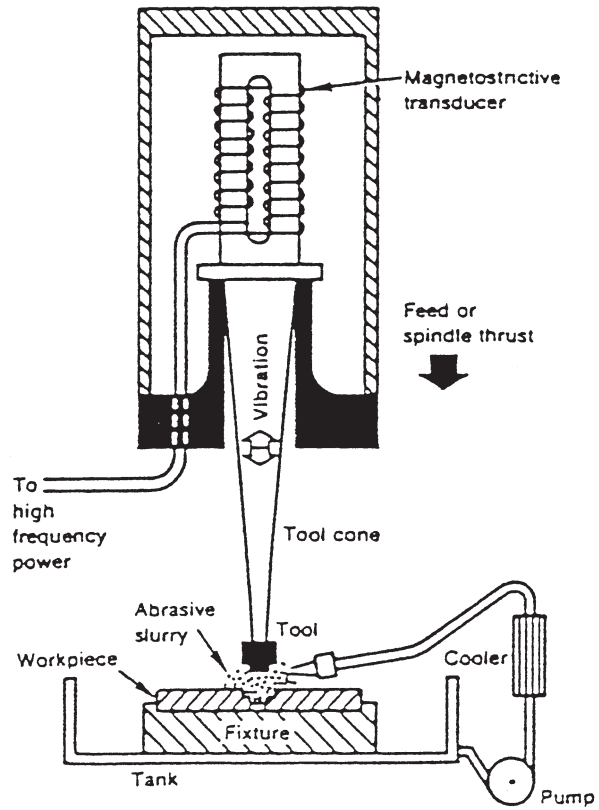


Figure 37 Ultrasonic machining.

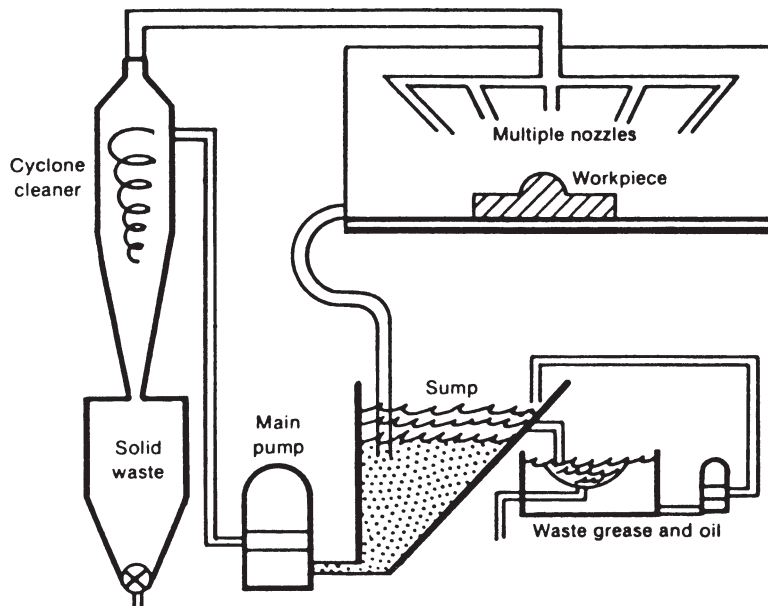


Figure 38 Water-jet machining.

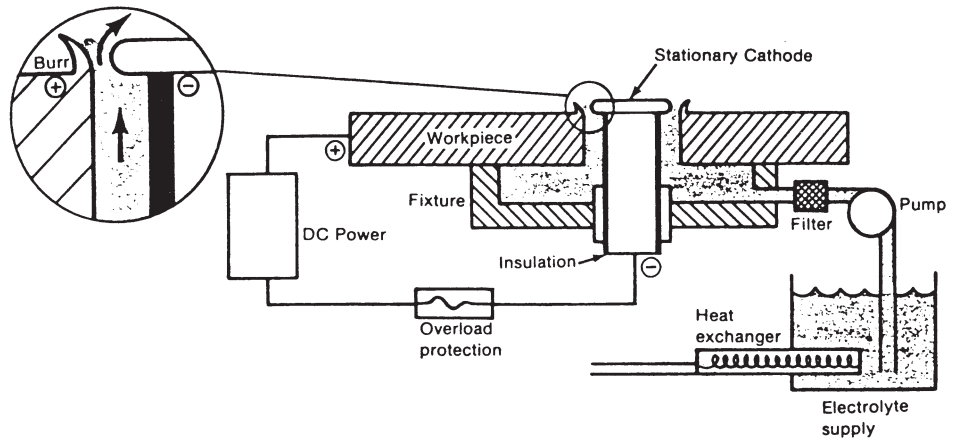


Figure 39 Electrochemical deburring.

16.10 Electrochemical Deburring

Electrochemical deburring (ECD) is a special version of ECM (Fig. 39). ECD was developed to remove burrs and fins or to round sharp corners. Anodic dissolution occurs on the workpiece burrs in the presence of a closely placed cathodic tool whose configuration matches the burred edge. Normally, only a small portion of the cathode is electrically exposed, so a maximum concentration of the electrolytic action is attained. The electrolyte flow usually is arranged to carry away any burrs that may break loose from the workpiece during the cycle. Voltages are low, current densities are high, electrolyte flow rate is modest, and electrolyte types are similar to those used for ECM. The electrode (tool) is stationary, so equipment is simpler than that used for ECM. Cycle time is short for deburring. Longer cycle time produces a natural radiusing action.

16.11 Electrochemical Discharge Grinding

Electrochemical discharge grinding (ECDG) combines the features of both electrochemical and electrical discharge methods of material removal (Fig. 40). ECDG has the arrangement and electrolytes of electrochemical grinding (ECG) but uses a graphite wheel without abrasive

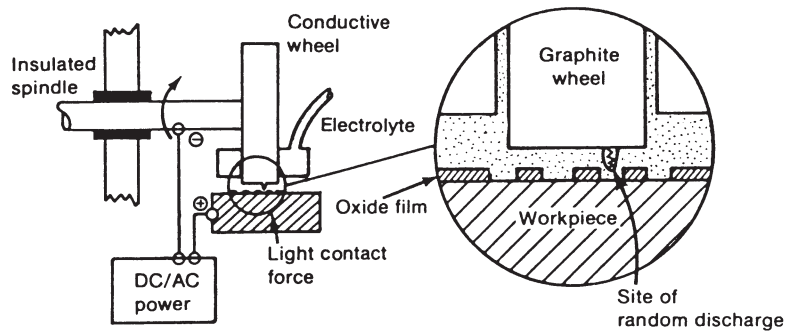


Figure 40 Electrochemical discharge grinding.

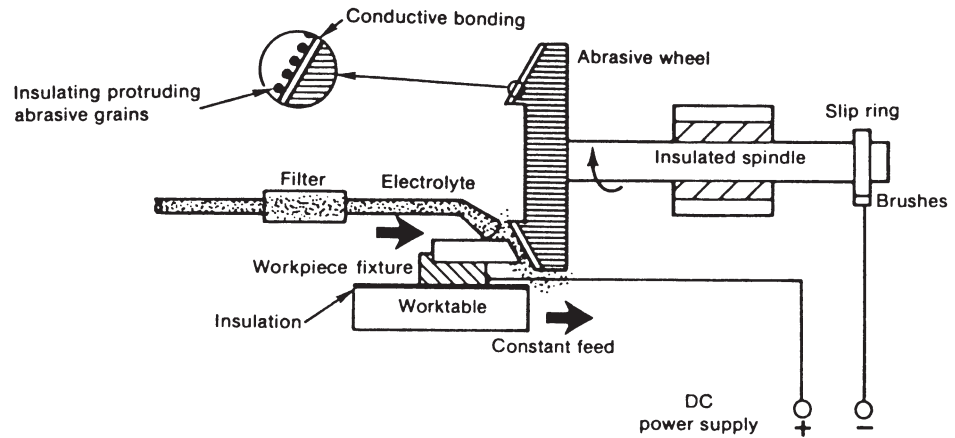


Figure 41 Electrochemical grinding.

grains. The random spark discharge is generated through the insulating oxide film on the workpiece by the power generated in an alternating-current (ac) source or by a pulsating direct-current (dc) source. The principal material removal comes from the electrolytic action of the low-level dc voltages. The spark discharges erode the anodic films to allow the electrolytic action to continue.

16.12 Electrochemical Grinding

Electrochemical grinding is a special form of electrochemical machining in which the conductive workpiece material is dissolved by anodic action, and any resulting films are removed by a rotating, conductive, abrasive wheel (Fig. 41). The abrasive grains protruding from the wheel form the insulating electrical gap between the wheel and the workpiece. This gap must be filled with electrolyte at all times. The conductive wheel uses conventional abrasives—aluminum oxide (because it is nonconductive) or diamond (for intricate shapes)—but lasts substantially longer than wheels used in conventional grinding. The reason for this is that the bulk of material removal (95–98%) occurs by deplating, while only a small amount (2–5%) occurs by abrasive mechanical action. Maximum wheel contact arc lengths are about $\frac{3}{4}$ –1 in. (19–25 mm) to prevent overheating the electrolyte. The fastest material removal is obtained by using the highest attainable current densities without boiling the electrolyte. The corrosive salts used as electrolytes should be filtered and flow rate should be controlled for the best process control.

16.13 Electrochemical Honing

Electrochemical honing (ECH) is the removal of material by anodic dissolution combined with mechanical abrasion from a rotating and reciprocating abrasive stone (carried on a spindle, which is the cathode) separated from the workpiece by a rapidly flowing electrolyte (Fig. 42). The principal material removal action comes from electrolytic dissolution. The abrasive stones are used to maintain size and to clean the surfaces to expose fresh metal to the electrolyte action. The small electrical gap is maintained by the nonconducting stones that are bonded to the expandable arbor with cement. The cement must be compatible with the electrolyte and the low dc voltage. The mechanical honing action uses materials, speeds, and pressures typical of conventional honing.

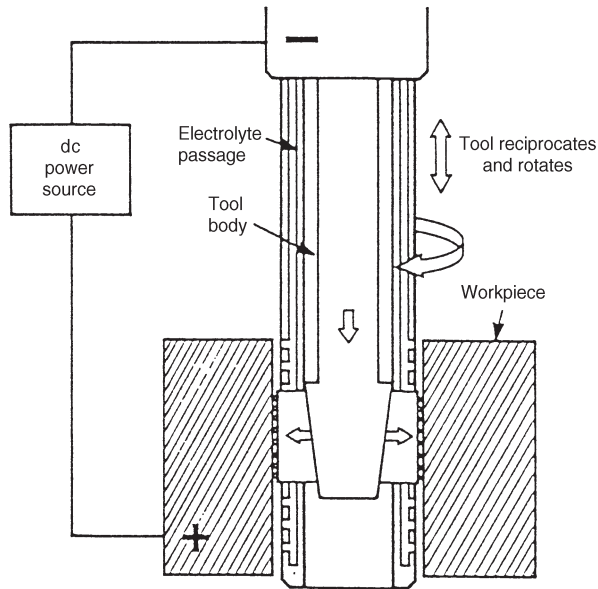


Figure 42 Electrochemical honing.

16.14 Electrochemical Machining

Electrochemical machining (ECM) is the removal of electrically conductive material by anodic dissolution in a rapidly flowing electrolyte, which separates the workpiece from a shaped electrode (Fig. 43). The filtered electrolyte is pumped under pressure and at controlled temperature to bring a controlled-conductivity fluid into the narrow gap of the cutting area. The shape imposed on the workpiece is nearly a mirror or conjugate image of the shape of the cathodic electrode. The electrode is advanced into the workpiece at a constant feed rate that exactly

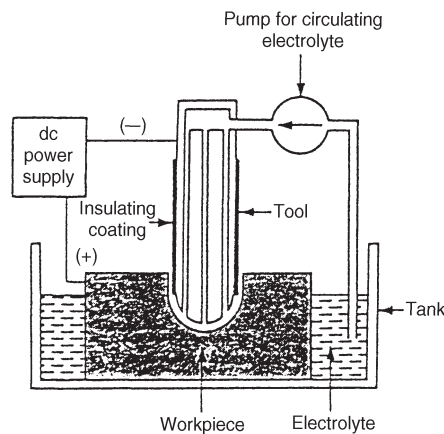


Figure 43 Electrochemical machining.

matches the rate of dissolution of the work material. Electrochemical machining is basically the reverse of electroplating.

Calculation of Metal Removal and Feed Rates in ECM

$$\text{Current } I = \frac{V}{R} \text{ amp}$$

$$\text{Resistance } R = \frac{g \times r}{A}$$

where g = length of gap, cm
 r = electrolyte resistivity
 A = area of current path, cm^2
 V = voltage
 R = resistance

$$\text{Current density } S = \frac{I}{A} = \frac{V}{r \times g} \text{ amp/cm}^2$$

The amount of material deposited or dissolved is proportional to the quantity of electricity passed (current \times time):

$$\text{Amount of material} = C \times I \times t$$

where C = constant
 t = time, s

The amount removed or deposited by one faraday (96,500 coulombs = 96,500 amp-s) is 1 gram-equivalent weight (G):

$$G = \frac{N}{n} \text{ (for 1 faraday)}$$

where N = atomic weight
 n = valence

$$\text{Volume of metal removed} = \frac{I \times t}{96,500} \times \frac{N}{n} \times \frac{1}{d} \times h$$

where d = density, g/cm^3
 h = current efficiency

$$\text{Specific removal rate } s = \frac{N}{n} \times \frac{1}{96,500} \times h \text{ cm}^3/\text{amp-s}$$

$$\text{Cathode feed rate } F = S \times s \text{ cm/s}$$

16.15 Electrochemical Polishing

Electrochemical polishing (ECP) is a special form of electrochemical machining arranged for cutting or polishing a workpiece (Fig. 44). Polishing parameters are similar in range to those for cutting but without the feed motion. ECP generally uses a larger gap and a lower current density than does ECM. This requires modestly higher voltages. [In contrast, electropolishing (ELP) uses still lower current densities, lower electrolyte flow, and more remote electrodes.]

16.16 Electrochemical Sharpening

Electrochemical sharpening (ECS) is a special form of electrochemical machining arranged to accomplish sharpening or polishing by hand (Fig. 45). A portable power pack and electrolyte

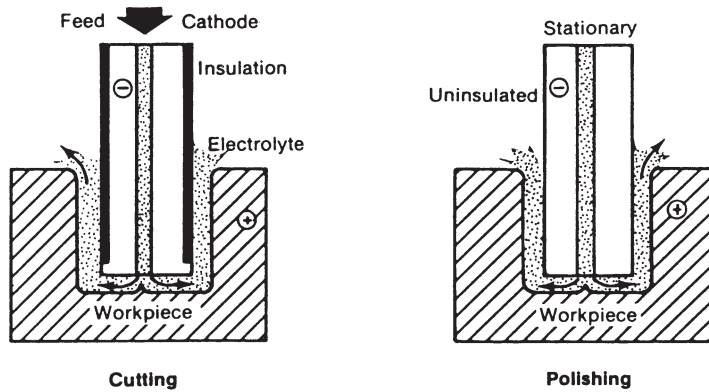


Figure 44 Electrochemical polishing.

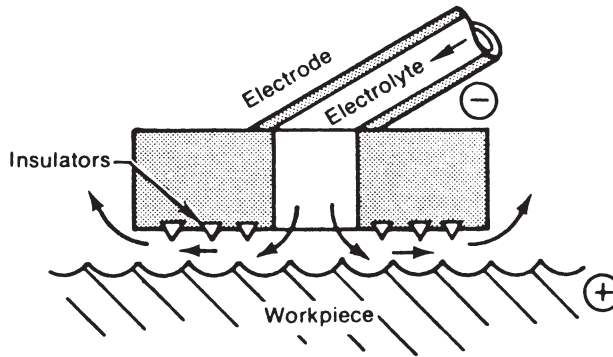


Figure 45 Electrochemical sharpening.

reservoir supply a finger-held electrode with a small current and flow. The fixed gap incorporated on the several styles of shaped electrodes controls the flow rate. A suction tube picks up the used electrolyte for recirculation after filtration.

16.17 Electrochemical Turning

Electrochemical turning (ECT) is a special form of electrochemical machining designed to accommodate rotating workpieces (Fig. 46). The rotation provides additional accuracy but complicates the equipment with the method of introducing the high currents to the rotating part. Electrolyte control may also be complicated because rotating seals are needed to direct the flow properly. Otherwise, the parameters and considerations of electrochemical machining apply equally to the turning mode.

16.18 Electrostream

Electrostream (ES) is a special version of electrochemical machining adapted for drilling very small holes using high voltages and acid electrolytes (see Fig. 47). The voltages are more than 10 times those employed in ECM or STEM, so special provisions for containment and

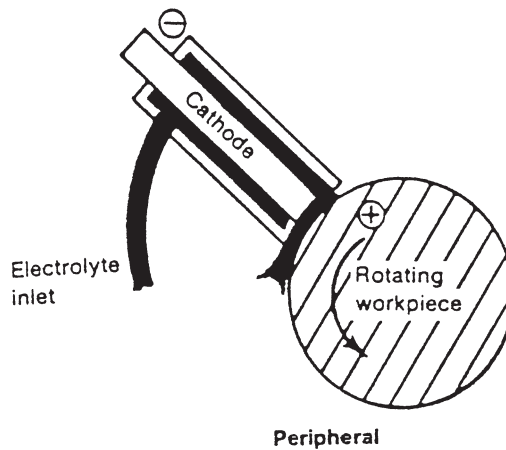


Figure 46 Electrochemical turning.

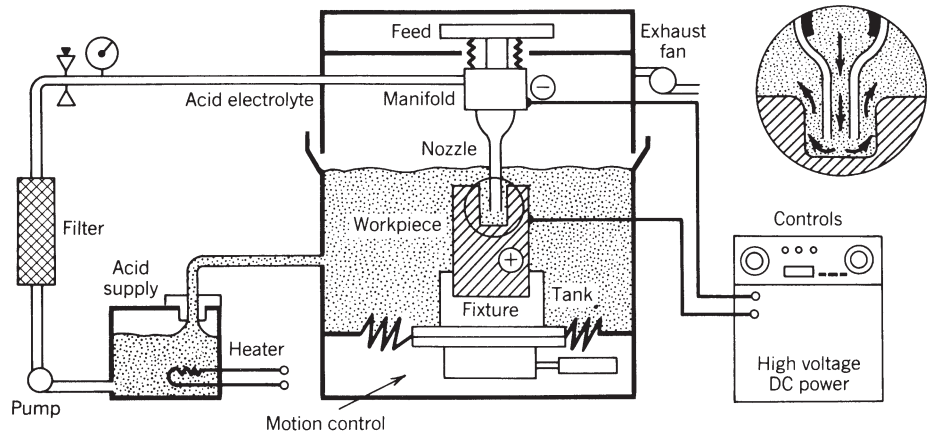


Figure 47 Electrostream.

protection are required. The tool is a drawn-glass nozzle, 0.001–0.002 in. smaller than the desired hole size. An electrode inside the nozzle or the manifold ensures electrical contact with the acid. Multiple-hole drilling is achieved successfully by ES.

16.19 Shaped-Tube Electrolytic Machining

Shaped-tube electrolytic machining (STEM) is a specialized ECM technique for “drilling” small, deep holes by using acid electrolytes (Fig. 48). Acid is used so that the dissolved metal will go into the solution rather than form a sludge, as is the case with the salt-type electrolytes of ECM. The electrode is a carefully straightened acid-resistant metal tube. The tube is coated with a film of enamel-type insulation. The acid is pressure-fed through the tube and returns via a narrow gap between the tube insulation and the hole wall. The electrode is fed into the workpiece at a rate exactly equal to the rate at which the workpiece material is dissolved. Multiple electrodes, even of varying diameters or shapes, may be used simultaneously. A solution

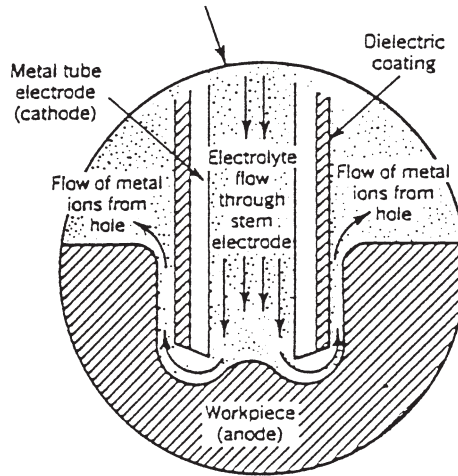


Figure 48 Shaped-tube electrolytic machining.

of sulfuric acid is frequently used as the electrolyte when machining nickel alloys. The electrolyte is heated and filtered, and flow monitors control the pressure. Tooling is frequently made of plastics, ceramics, or titanium alloys to withstand the electrified hot acid.

16.20 Electron Beam Machining

Electron beam machining (EBM) removes material by melting and vaporizing the workpiece at the point of impingement of a focused stream of high-velocity electrons (Fig. 49). To eliminate scattering of the beam of electrons by contact with gas molecules, the work is done in a high-vacuum chamber. Electrons emanate from a triode electron-beam gun and are accelerated

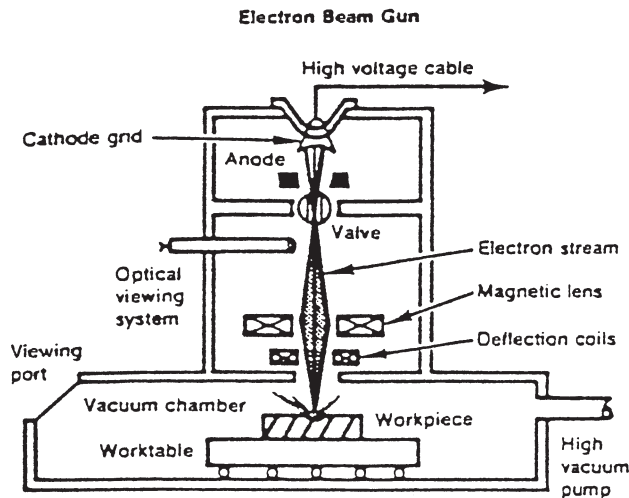


Figure 49 Electron beam machining.

to three-fourths the speed of light at the anode. The collision of the electrons with the workpiece immediately translates their kinetic energy into thermal energy. The low-inertia beam can be simply controlled by electromagnetic fields. Magnetic lenses focus the electron beam on the workpiece, where a 0.001-in. (0.025-mm) diameter spot can attain an energy density of up to 10^9 W/in.² (1.55×10^8 W/cm²) to melt and vaporize any material. The extremely fast response time of the beam is an excellent companion for three-dimensional computer control of beam deflection, beam focus, beam intensity, and workpiece motion.

16.21 Electrical Discharge Grinding

Electrical discharge grinding (EDG) is the removal of a conductive material by rapid, repetitive spark discharges between a rotating tool and the workpiece, which are separated by a flowing dielectric fluid (Fig. 50). (EDG is similar to EDM except that the electrode is in the form of a grinding wheel and the current is usually lower.) The spark gap is servocontrolled. The insulated wheel and the worktable are connected to the dc pulse generator. Higher currents produce faster cutting, rougher finishes, and deeper heat-affected zones in the workpiece.

16.22 Electrical Discharge Machining

Electrical discharge machining (EDM) removes electrically conductive material by means of rapid, repetitive spark discharges from a pulsating dc power supply with dielectric flowing between the workpiece and the tool (Fig. 51). The cutting tool (electrode) is made of electrically conductive material, usually carbon. The shaped tool is fed into the workpiece under servocontrol. A spark discharge then breaks down the dielectric fluid. The frequency and energy per spark are set and controlled with a dc power source. The servocontrol maintains a constant gap between the tool and the workpiece while advancing the electrode. The dielectric oil cools and flushes out the vaporized and condensed material while reestablishing insulation in the gap. Material removal rate ranges from 16 to 245 cm³/h. EDM is suitable for cutting materials regardless of their hardness or toughness. Round or irregular-shaped holes 0.002 in. (0.05 mm) diameter can be produced with L/D ratio of 20:1. Narrow slots as small as 0.002–0.010 in. (0.05–0.25 mm) wide are cut by EDM.

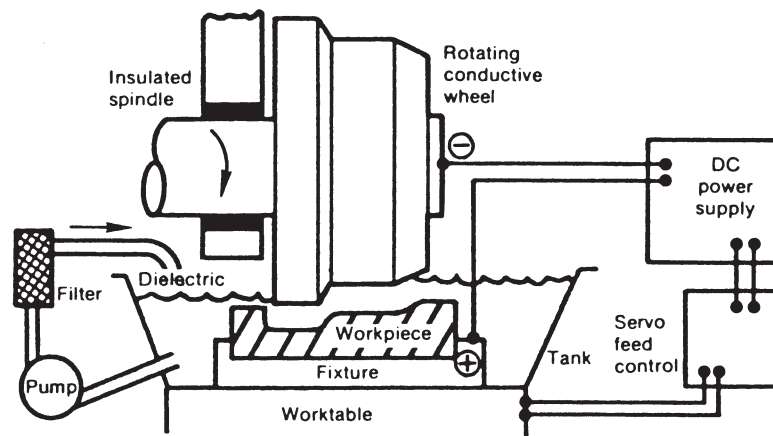


Figure 50 Electrical discharge grinding.

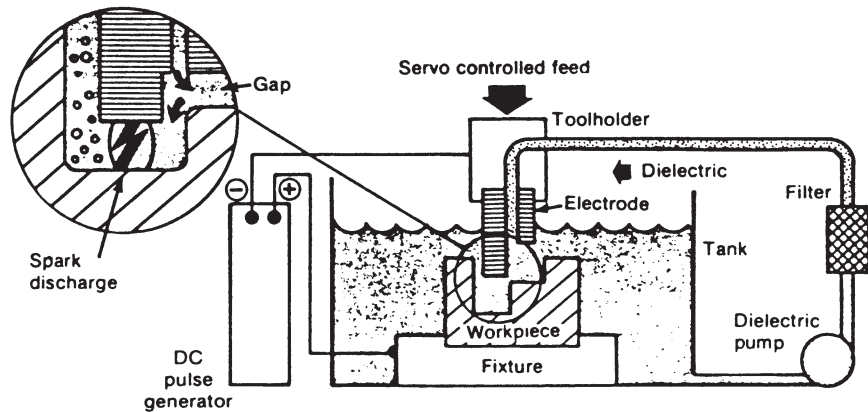


Figure 51 Electrical discharge machining.

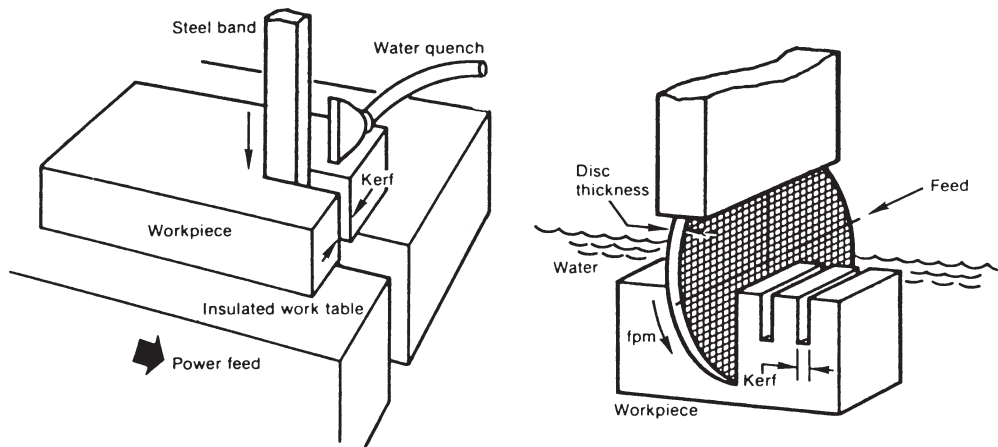


Figure 52 Electrical discharge sawing.

16.23 Electrical Discharge Sawing

Electrical discharge sawing (EDS) is a variation of EDM that combines the motion of either a band saw or a circular disk saw with electrical erosion of the workpiece (Fig. 52). The rapid-moving, untoothed, thin, special steel band or disk is guided into the workpiece by carbide-faced inserts. A kerf only 0.002–0.005 in. (0.050–0.13 mm) wider than the blade or disk is formed as they are fed into the workpiece. Water is used as a cooling quenchant for the tool, swarf, and workpiece. Circular cutting is usually performed under water, thereby reducing noise and fumes. While the work is power fed into the band (or the disk into the work), it is not subjected to appreciable forces because the arc does the cutting, so fixturing can be minimal.

16.24 Electrical Discharge Wire Cutting (Traveling Wire)

Electrical discharge wire cutting (EDWC) is a special form of electrical discharge machining wherein the electrode is a continuously moving conductive wire (Fig. 53). EDWC is often

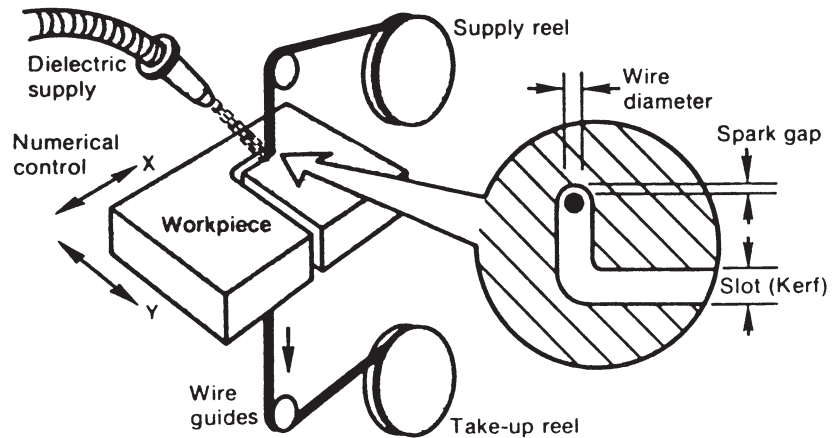


Figure 53 Electrical discharge wire cutting.

called *traveling wire* EDM. A small-diameter tension wire, 0.001–0.012 in. (0.03–0.30 mm), is guided to produce a straight, narrow-kerf size 0.003–0.015 in. (0.075–0.375 mm). Usually, a programmed or numerically controlled motion guides the cutting, while the width of the kerf is maintained by the wire size and discharge controls. The dielectric is oil or deionized water carried into the gap by motion of the wire. Wire EDM is able to cut plates as thick as 12 in. (300 mm) and issued for making dies from hard metals. The wire travels with speed in the range of 6–300 in./min (0.15–8 mm/min). A typical cutting rate is 1 in.² (645 mm²) of cross-sectional area per hour.

16.25 Laser Beam Machining

Laser beam machining (LBM) removes material by melting, ablating, and vaporizing the workpiece at the point of impingement of a highly focused beam of coherent monochromatic light (Fig. 54). Laser is an acronym for “light amplification by stimulated emission of radiation.”

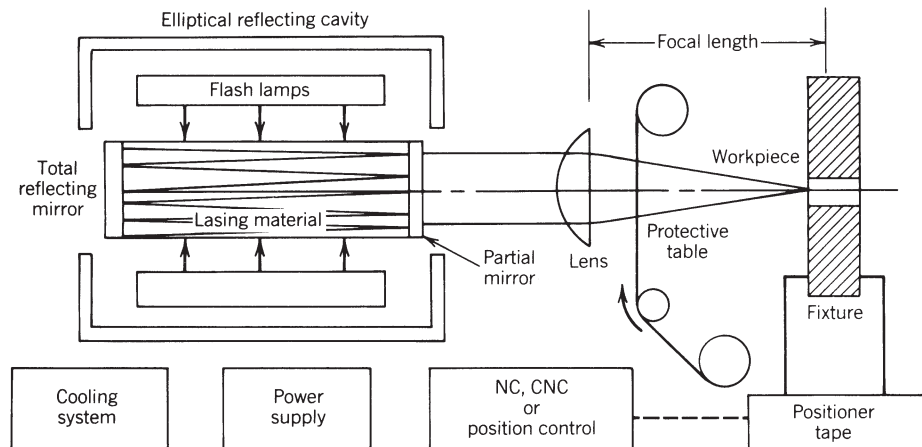


Figure 54 Laser beam machining.

The electromagnetic radiation operates at wavelengths from the visible to the infrared. The principal lasers used for material removal are the Nd:glass (neodymium:glass), the Nd:YAG (neodymium:yttrium–aluminum–garnet), and the ruby and the carbon dioxide (CO₂). The last is a gas laser (most frequently used as a torch with an assisting gas—see Section 6.26), while others are solid-state lasing materials.

For pulsed operation, the power supply produces short, intense bursts of electricity into the flash lamps, which concentrate their light flux on the lasing material. The resulting energy from the excited atoms is released at a characteristic, constant frequency. The monochromatic light is amplified during successive reflections from the mirrors. The thoroughly collimated light exits through the partially reflecting mirror to the lens, which focuses it on or just below the surface of the workpiece. The small beam divergence, high peak power, and single frequency provide excellent, small-diameter spots of light with energy densities up to 3×10^{10} W/in.² (4.6×10^9 W/cm²), which can sublime almost any material. Cutting requires energy densities of 10^7 – 10^9 W/in.² (1.55×10^6 – 1.55×10^8 W/cm²), at which rate the thermal capacity of most materials cannot conduct energy into the body of the workpiece fast enough to prevent melting and vaporization. Some lasers can instantaneously produce 41,000°C (74,000°F). Holes of 0.001 in. (0.025 mm), with depth-to-diameter 50 to 1 are typically produced in various materials by LBM.

16.26 Laser Beam Torch

Laser beam torch (LBT) is a process in which material is removed by the simultaneous focusing of a laser beam and a gas stream on the workpiece (see Fig. 55). A continuous-wave (CW) laser or a pulsed laser with more than 100 pulses per second is focused on or slightly below the surface of the workpiece, and the absorbed energy causes localized melting. An oxygen gas stream promotes an exothermic reaction and purges the molten material from the cut. Argon or nitrogen gas is sometimes used to purge the molten material while also protecting the workpiece.

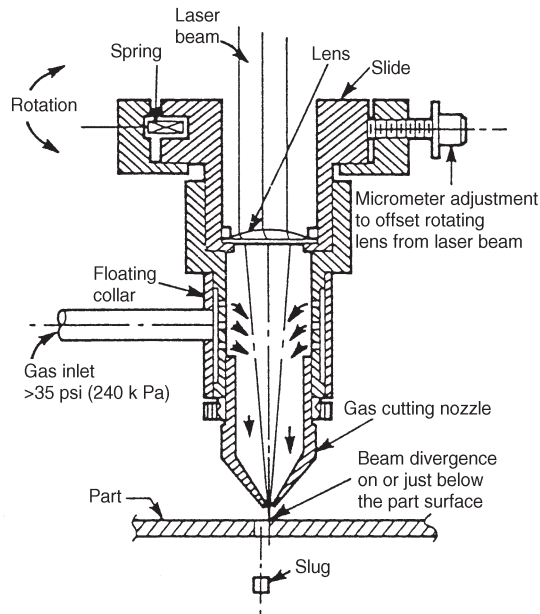


Figure 55 Laser beam torch.

Argon or nitrogen gas is often used when organic or ceramic materials are being cut. Close control of the spot size and the focus on the workpiece surface is required for uniform cutting. The type of gas used has only a modest effect on laser penetrating ability. Typically, short laser pulses with high peak power are used for cutting and welding. The CO₂ laser is the laser most often used for cutting. Thin materials are cut at high rates, $\frac{1}{8}$ – $\frac{3}{8}$ in. (3.2–9.5 mm) thickness is a practical limit.

16.27 Plasma Beam Machining

Plasma beam machining (PBM) removes material by using a superheated stream of electrically ionized gas (Fig. 56). The 20,000–50,000°F (11,000–28,000°C) plasma is created inside a water-cooled nozzle by electrically ionizing a suitable gas, such as nitrogen, hydrogen, or argon, or mixtures of these gases. Since the process does not rely on the heat of combustion between the gas and the workpiece material, it can be used on almost any conductive metal. Generally, the arc is transferred to the workpiece, which is made electrically positive. The plasma—a mixture of free electrons, positively charged ions, and neutral atoms—is initiated in a confined, gas-filled chamber by a high-frequency spark. The high-voltage dc power sustains the arc, which exits from the nozzle at near-sonic velocity. The high-velocity gases blow away the molten metal “chips.” Dual-flow torches use a secondary gas or water shield to assist in blowing the molten metal out of the kerf, giving a cleaner cut. PBM is sometimes called

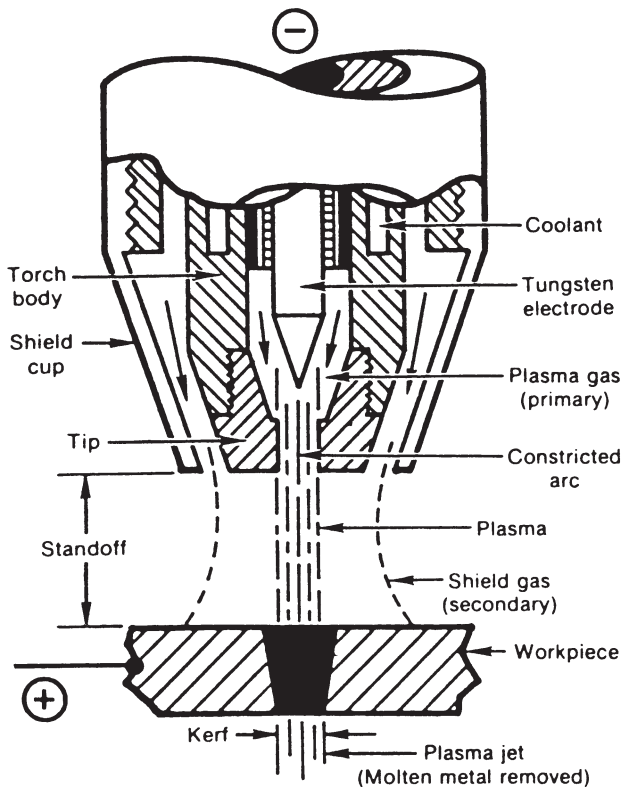


Figure 56 Plasma beam machining.

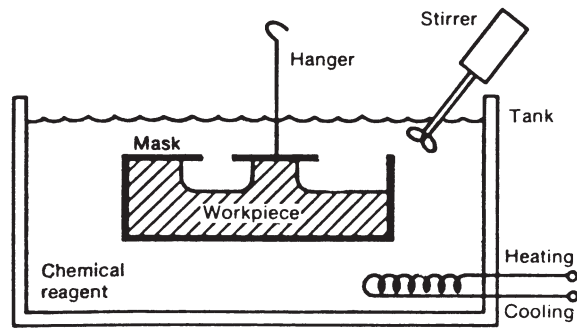


Figure 57 Chemical machining.

plasma-arc cutting (PAC). PBM can cut plates up to 6.0 in. (152 mm) thick. Kerf width can be as small as 0.06 in. (1.52 mm) in cutting thin plates.

16.28 Chemical Machining: Chemical Milling, Chemical Blanking

Chemical machining (CHM) is the controlled dissolution of a workpiece material by contact with a strong chemical reagent (Fig. 57). The thoroughly cleaned workpiece is covered with a strippable, chemically resistant mask. Areas where chemical action is desired are outlined on the workpiece with the use of a template and then stripped off the mask. The workpiece is then submerged in the chemical reagent to remove material simultaneously from all exposed surfaces. The solution should be stirred or the workpiece should be agitated for more effective and more uniform action. Increasing the temperatures will also expedite the action. The machined workpiece is then washed and rinsed, and the remaining mask is removed. Multiple parts can be maintained simultaneously in the same tank. A wide variety of metals can be chemically machined; however, the practical limitations for depth of cut are 0.25–0.5 in. (6.0–12.0 mm) and typical etching rate is 0.001 in./min (0.025 mm/min).

In chemical blanking, the material is removed by chemical dissolution instead of shearing. The operation is applicable to production of complex shapes in thin sheets of metal.

16.29 Electropolishing

Electropolishing (ELP) is a specialized form of chemical machining that uses an electrical deplating action to enhance the chemical action (Fig. 58). The chemical action from the concentrated heavy acids does most of the work, while the electrical action smooths or polishes the irregularities. A metal cathode is connected to a low-voltage, low-amperage dc power source and is installed in the chemical bath near the workpiece. Usually, the cathode is not shaped or conformed to the surface being polished. The cutting action takes place over the entire exposed surface; therefore, a good flow of heated, fresh chemicals is needed in the cutting area to secure uniform finishes. The cutting action will concentrate first on burrs, fins, and sharp corners. Masking, similar to that used with CHM, prevents cutting in unwanted areas. Typical roughness values range from 4 to 32 $\mu\text{in.}$ (0.1–0.8 μm).

16.30 Photochemical Machining

Photochemical machining (PCM) is a variation of CHM where the chemically resistant mask is applied to the workpiece by a photographic technique (Fig. 59). A photographic negative,

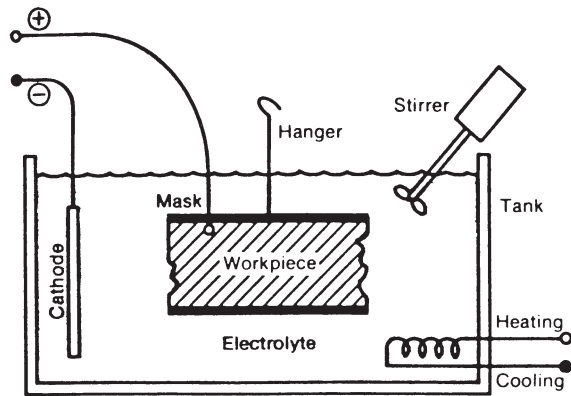


Figure 58 Electropolishing.

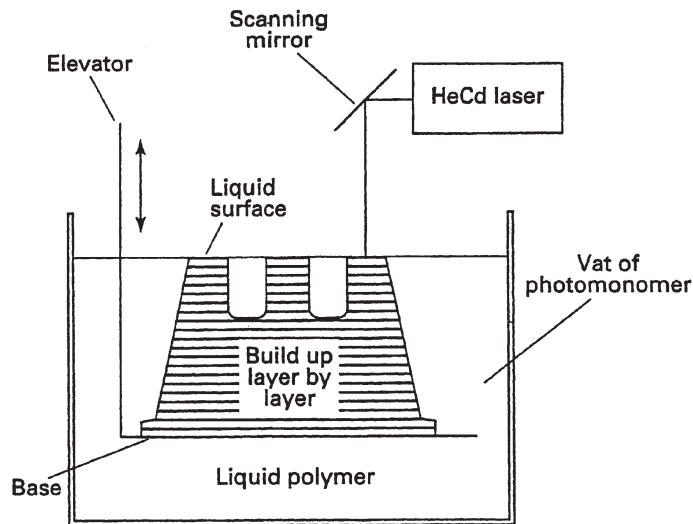


Figure 59 Rapid prototyping using laser to photopolymerize the liquid photopolymer.

often a reduced image of an oversize master print (up to 100 \times), is applied to the workpiece and developed. Precise registry of duplicate negatives on each side of the sheet is essential for accurately blanked parts. Immersion or spray etching is used to remove the exposed material. The chemicals used must be active on the workpiece, but inactive against the photoresistant mask. The use of PCM is limited to thin materials—up to $1/16$ in. (1.5 mm).

16.31 Thermochemical Machining

Thermochemical machining (TCM) removes the workpiece material—usually only burrs and fins—by exposure of the workpiece to hot, corrosive gases. The process is sometimes called *combustion machining*, *thermal deburring*, or *thermal energy method* (TEM). The workpiece is exposed for a very short time to extremely hot gases, which are formed by detonating

an explosive mixture. The ignition of the explosive—usually hydrogen or natural gas and oxygen—creates a transient thermal wave that vaporizes the burrs and fins. The main body of the workpiece remains unaffected and relatively cool because of its low surface-to-mass ratio and the shortness of the exposure to high temperatures.

16.32 Rapid Prototyping and Rapid Tooling

In the past, when making a prototype, a full-scale model of a product, the designed part would have then machined or sculptured from wood, plastic, metal, or other solid materials. Now there is rapid prototyping (Fig. 59), also called desktop manufacturing, a process by which a solid physical model of a product is made directly from a three-dimensional computer-aided design (CAD) drawing.

Rapid prototyping entails several different consolidation techniques and steps: resin curing, deposition, solidification, and finishing. The conceptual design is viewed in its entirety and at different angles on the monitor through a three-dimensional CAD system. The part is then sliced into horizontal planes from 0.004 to 0.008 in. (0.10–0.20 mm). Then a helium:cadmium (He:Cd) laser beam passes over the liquid photopolymer resin. The ultraviolet (UV) photons harden the photosensitive resin. The part is lowered only one layer thickness. The recoater blade sweeps over the previously hardened surface, applying a thin, even coat of resin. Upon completion, a high-intensity broadband or continuum ultraviolet radiation is used to cure the mold. Large parts can be produced in sections, and then the sections are welded together.

Other techniques, such as selective laser sintering (SLS) (Fig. 60), use a thin layer of heat-fusible powder that has been evenly deposited by a roller. A CO₂ laser, controlled by a CAD program, heats the powder to just below the melting point and fuses it only along the programmed path.

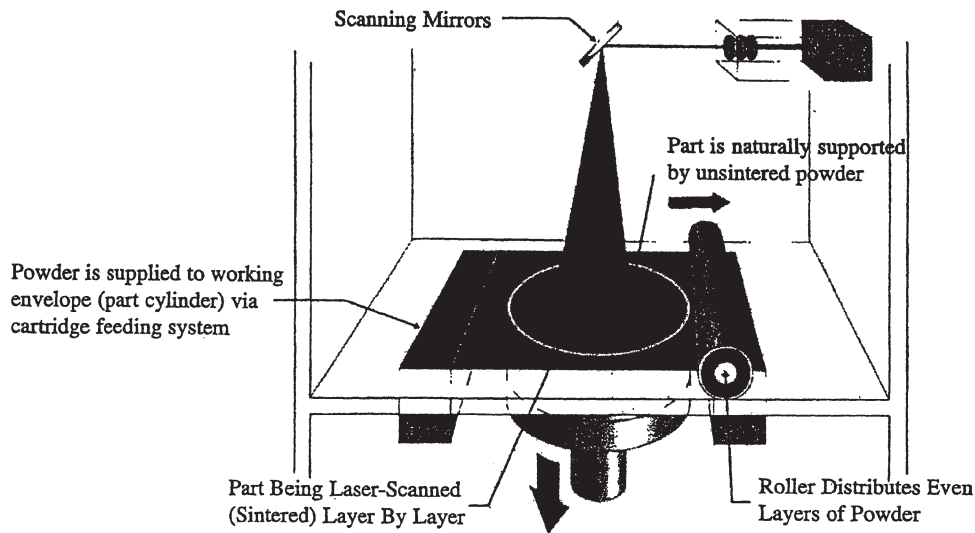


Figure 60 Rapid prototyping using sintering process (powder).

REFERENCES

1. Society of Manufacturing Engineers, *Tool and Manufacturing Engineers Handbook*, Vol. 1, Machining, McGraw-Hill, New York, 1985.
2. *Machining Data Handbook*, 3rd ed., Machinability Data Center, Cincinnati, OH, 1980.
3. *Metals Handbook*, 8th ed., Vol. 3, Machining American Society for Metals, Metals Park, OH, 1985.
4. R. LeGrand (Ed.), *American Machinist's Handbook*, 3rd ed., McGraw-Hill, New York, 1973.
5. *Machinery's Handbook*, 21st ed., Industrial Press, New York, 1979.
6. *Machinery Handbook*, Vol. 2, Machinability Data Center, Department of Defense, Cincinnati, OH, 1983.
7. K. G. Swift and J. D. Booker, *Process Selection*, Arnold, London, 1977.
8. C. Sommer, *Non-traditional Machining Handbook*, Advance Publishing, Houston, 2000.

BIBLIOGRAPHY

- B. H., Amstead, P. F. Ostwald, and M. L. Begeman, *Manufacturing Processes*, 8th ed., Wiley, New York, 1988.
- ASM Handbook*, Vol. 16: Machining, ASM International, Materials Park, OH, 1995.
- V. P., Astakhov, *Metal Cutting Mechanics*, CRC Press, Boca Raton, FL, 1998.
- J., Brown, *Advanced Machining Technology Handbook*, McGraw-Hill, New York, 1998.
- J. A., Charles, F. Crane, and J. Furness, *Selection and Use of Engineering Materials*, 3rd ed., Butterworth Heinemann, England, 1997.
- E. P., DeGarmo, J. T. Black, and R. A. Kohser, *Material and Processes in Manufacturing*, 9th ed., Wiley, Hoboken, NJ, 2003.
- L. E., Doyle, G. F. Schrader, and M. B. Singer, *Manufacturing Processes and Materials for Engineers*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ, 1985.
- S. D., El Walkil, *Processes and Design for Manufacturing*, 2nd ed., PWS, Boston, 1998.
- M. P., Groover, *Automation, Production Systems and Computer-Integrated Manufacturing*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2001.
- B. J., Hamrock, B. Jacobson, and S. R. Schmid, *Fundamentals of Machine Elements*, McGraw-Hill, Boston, 1999.
- V. K., Jain and P. C. Pandey, *Theory and Practice of Electro-chemical Machining*, Wiley, New York, 1993.
- S., Kalpakjia and S. R. Schmid, *Manufacturing Processes for Engineering Materials*, Prentice-Hall, Englewood Cliffs, NJ, 2003.
- M., Kronenberg, *Machining Science and Application*, Pergamon, London, 1966.
- R. A., Lindberg, *Processes and Materials of Manufacture*, 2nd ed., Allyn & Bacon, Boston, 1977.
- J. A., McGeough, *Advanced Methods of Machining*, Wolters Kluwer, Dordrecht, The Netherlands, 1988.
- Metal Cutting Tool Handbook*, 7th ed., Industrial Press, Cleveland, OH, 1989.
- H. D., Moore, and D. R. Kibbey, *Manufacturing Materials and Processes*, 3rd ed., Wiley, New York, 1982.
- B. W., Niebel, and A. B. Draper, *Product Design and Process Engineering*, McGraw-Hill, New York, 1974.
- J. A., Schey, *Introduction to Manufacturing Processes*, McGraw-Hill, New York, 1977.
- M. C., Shaw, *Metal Cutting Principles*, Oxford University Press, Oxford, UK, 1984.
- C., Sommer, *Non-traditional Machining Handbook*, Advance Publishing, Houston, TX, 2000.
- E. M., Trent, and P. K. Wright, *Metal Cutting*, 4th ed., Butterworth Heinemann, England, 1999.
- R. A., Walsh, *Machining and Metalworking Handbook*, McGraw-Hill, New York, 1994.
- F., Waters, *Fundamentals of Manufacturing for Engineers*, UCL Press, University College, London, 1996.
- J. A., Webster, *Abrasive Processes Theory, Technology, and Practice*, Dekker, New York, 1996.
- M. E., Zohdi, "Statistical Analysis, Estimation and Optimization in the Grinding Process," *ASME Trans.*, Paper No. 73-DET-3, 1973.

CHAPTER 5

MANUFACTURING SYSTEMS EVALUATION

Walter W. Olson
University of Toledo
Toledo, Ohio

1 INTRODUCTION	183	5 ASSESSMENT	189
2 COMPONENTS OF ECM	184	5.1 Assessment Planning	189
3 MANUFACTURING SYSTEMS	185	5.2 Data Collection	190
3.1 Levels of Manufacturing Systems	185	5.3 Site Visit and Inspection	191
3.2 Plan–Do–Check–Act Cycle	187	5.4 Reporting and Project Formulation	192
4 SYSTEM EFFECTS ON ECM	187	6 SUMMARY	193
		REFERENCES	193

1 INTRODUCTION

Environmentally conscious manufacturing (ECM) is the production of products using processes and techniques selected to both be economically viable and have the least impact on the environment. Process selection criteria include minimum waste, minimum use of hazardous materials, and minimum use of energy in addition to the production goals. Products produced in this manner are often more competitive in the marketplace because these criteria result in cost reduction, cost avoidance, and increased appeal to the consumer.

This chapter discusses several techniques for assessing and evaluating ECM performance of manufacturing systems. These techniques begin with an analysis of the paperwork/data followed by a visit of the plant floor and conducting interviews. This provides the basis for making evaluations of improvement areas and performing the tasks necessary to formulate improvement projects. The emphasis here is on improving the manufacturing systems; however, the benefits occur in production.

This chapter focuses on very practical and fundamental issues in manufacturing. As a result, it is not readily applicable to a firm seeking International Organization for Standardization (ISO) 14001 accreditation. Nor is it meant to be. Whereas ISO 14001 is an environmental management system, this chapter is about manufacturing systems and their evaluation. Ghisellinia and Thurston identified several decision traps that ISO 14001 imposes “the ‘management’ nature of the standard, failure to identify a rigorous environmental baseline, misconception of pollution prevention, inordinate emphasis on short-term goals, focusing on regulatory compliance, and diversion of EMS resources to the documentation system.”¹ The processes here seek to avoid these traps. Therefore, there will be little reference to ISO 14001 or life-cycle assessment (LCA) in the following pages.

Manufacturing systems are the planning, communications, coordination, monitoring, and management aspects of manufacturing. Industrial engineers often perform these tasks within the overall manufacturing system. The goal of this chapter is to provide guidance to assist the

industrial engineer in finding better methods and techniques to make the overall manufacturing system more responsive to the goals of ECM.

2 COMPONENTS OF ECM

Although one can reduce ECM in a number of specific details, all of which seem independent of each other, the engineer would do well to remember three major points:

1. Reduce waste.
2. Reduce hazardous materials and processes.
3. Reduce energy.

Almost every other ECM detail relates to these simple three objectives.

Waste is probably the single most important element. *Wastes are expenditures of resources that are not incorporated into the product.* Resources used in manufacturing are time, money, capital, labor, and materials. Wastes are characterized by emissions, machining offal, cutting fluids, person hours expended in waiting, buffering of items waiting for the next process, and machine downtime due to changeover or maintenance, for example. When waste occurs, the product is more expensive.

A major division of manufacturing processes results in material separations. It is important to realize that it may be necessary to have waste in order to produce a product (see Fig. 1). However, the material lost should be minimized. A systematic approach to identifying and eliminating waste results in increased ECM while also increasing profitability. Often trade-offs must be made. For example, to be economically viable a process may require the use of a more wasteful material than one where material waste can essentially be eliminated but at a cost that prevents economic production of the product. Although this is an extreme example that is rarely encountered, effective ECM requires that several objectives be met while producing a product that can be marketed successfully. The greatest waste occurs when a product was produced in quantity that failed to meet market requirements. Thus, it is important to identify wastes and know exactly where wastes are occurring, but it is even more important to know how to reduce waste without incurring greater costs.

Hazardous materials are defined as materials that, by their nature, chemistry, or conversion, result in reduced health of those exposed to them, degraded safety, or environmental risks produced by their uncontrolled release. These include toxic, mutagenic, radioactive, inflammable, and explosive substances. In addition to substances, hazards can include such things as noise, compressed fluids, and dusts. In many cases, specific controls and permissible levels are associated to these substances imposed by governments. Because of the additional

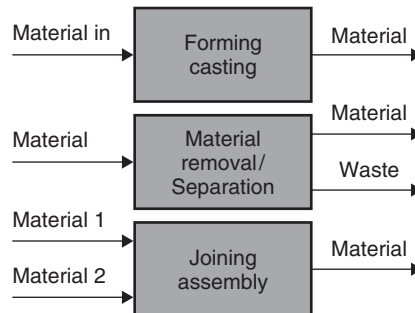


Figure 1 Classification of manufacturing processes.

controls and the additional increased handling awareness, use of these materials increases manufacturing costs.

Every industrial engineer in a manufacturing plant is intimately familiar with the cost of energy. However, they are not necessarily familiar with the environmental damage that energy production creates. Regardless of the form of energy or the source of energy, all energy has associated environmental damage. For example, direct solar conversion, which many vaunt as clean energy, requires the use of cadmium, selenium, and tellurium—all of which are harmful to the environment. Therefore, the reduction of energy requirements not only reduces manufacturing costs but also reduces environmental damage. Thus, reduction of energy is also an objective of ECM.

3 MANUFACTURING SYSTEMS

Systems are required to produce products. A system is an organized whole consisting of sub-systems that receive inputs from the external environment, transforming the inputs to outputs.² Many manufacturing activities required well-defined, systematic approaches to ensure efficiency. Manufacturing requires close coordination of labor, tools, materials, and information. If the systems that ensure this coordination are missing or flawed in their performance, waste is produced. These systems must start with the initial concept of the product and extend through the warranty of the product after it is in use. In some cases, this also includes the disposition of the product after the usage phase.

3.1 Levels of Manufacturing Systems

Manufacturing systems can be considered at four different levels of refinement (see Fig. 2). At the top level, loosely called the *metasystem*, are decisions about the organization, the subcomponents, and the interactions of the components within the overall manufacturing systems. Factors that influence the metasystem are the top decisions as to what the foci of the businesses are, whether to produce to order or produce to stock, the complexity of the

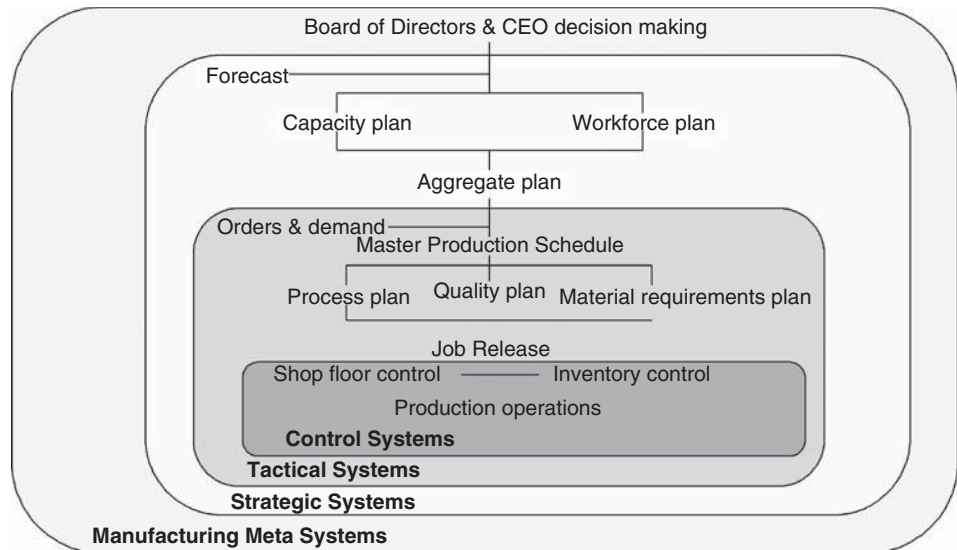


Figure 2 Manufacturing system levels.

product, and the economic responsiveness needed. These will determine the components of the manufacturing system and establish the outputs of the components. The planning horizon for these decisions is the expected life of the corporation. It ranges from years to decades.

At the *strategic* level, forecasting and capacity planning (including facility planning and workforce planning) are the major subsystems. They are taken together with an output of the aggregate plan. The aggregate plan is based on the capacity of the manufacturing firm and includes the labor policies to drive the resources at the firm level. Aggregate planning horizons typically range from three months to five years, depending on the size of the firm, the product complexity, and the commitment levels of resources needed. Since it might take two years or more to bring a new factory online, aggregate plans need to incorporate the long-term economic outlook of the corporation and be consonant with the corporation strategic plan. In fact, one could argue that the aggregate plan is the operational statement of the strategic plan.

The next level down is the *tactical* level of systems. Here, process plans, quality plans, and production scheduling need to be developed. Actual customer orders enter through a demand management system. Depending on decisions made at the metasytem level, orders are met from stock or by introduction of new production orders to the operations. Critical decisions at the tactical level determine what tools, what level of quality, and what response times are needed to meet the aggregate plan. The major output of the tactical systems is the production master schedule. At the tactical level, planning horizons range from one week to more than a year.

At the lowest system level are *control* systems. Shop floor control, inventory control, quality control, and maintenance control systems are the major components. These controls are driven by the production master schedule and direct and monitor resources to meet the production master schedule. Planning horizons may involve activities of a few seconds up to three months.

Decisions that have the most far-reaching effects occur at the metasytem level. As decisions are made at lower levels, they decrease in scope and reach but become more detailed. Thus, decisions at high levels involve more movement of resources and have greater potential for waste reduction than do decisions at the lowest levels.

Although various forms of manufacturing systems may have different names for the systems above and stipulate a set of interactions between the systems, any complete manufacturing system will have subsystems performing the functions just described. Essentially, these systems relate real-world factors to models that can be used to direct future manufacturing activities. In this context, it is essential that all systems implement a *plan–do–check–act (PDCA) cycle*. This is illustrated in Fig. 3.

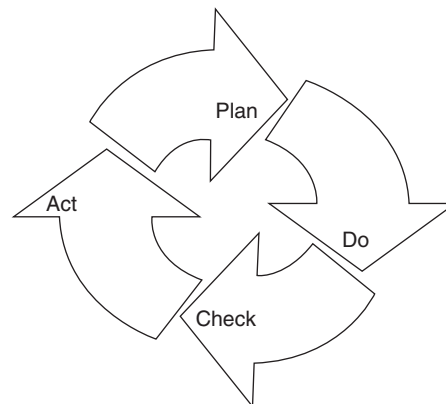


Figure 3 Plan–do–check–act cycle.

3.2 Plan–Do–Check–Act Cycle

The PDCA cycle, sometimes called the Deming cycle, expresses a relationship between planning, operations/production, evaluations of outcome, and management that are essential in a well-organized and profitable manufacturing system. The “plan” phase of the cycle prepares the system to meet its operational objectives and goals. In this phase, the events necessary for performance of the system are organized and the resources needed for production are scheduled. Planning is fed by the action phase, which expresses management intent. The “do” phase is the expression of the system performance. In most cases, the system is named after the do phase as this is where the system produces its product and outputs. The “check” phase is the assessment and evaluation of the do and plan phases. In the check phase, it is important to observe “what went right” as well as “what went wrong.” This is often the role of quality control within the system. Finally, the “act” phase is the role of management in the system. Leaders and managers set goals and objectives based on the reports from the check phase and their observations of the do and plan phases. Furthermore, they provide the directive action to correct and improve the system for the next cycle of PDCA.

Within manufacturing systems, the majority of methods to detect system wastes are based on measuring the system variability. In general, if a system is highly variable, the system is not performing well. The major causes of system variability are insufficient resources to perform the assigned tasks, insufficient capacity to handle the workload, and insufficient time for planning and execution. Oftentimes, these problems are exacerbated by failure of higher level systems that provide inputs to the system in question. For example, aggregate plan changes frequently raise havoc with production scheduling and material ordering systems in the attempt to keep the master production schedule responsive. As results, long-lead-time materials may be ordered, labor hired, and plant resources designated that are either too much (hence, wasteful) or too little (again, wasteful of time) when actual production is executed.

Overall, the manufacturing systems organize production in an orderly way from high-level goals to the very detailed and mundane activities. Inefficient manufacturing systems result in inefficient production operations. Thus, observations of inefficiency at the operational level must be tracked to the level of system where the inefficiency first was manifested. This process reveals improvement opportunities. Thus far, these systems have been considered independently of ECM. In the next section, it will be shown how these systems affect ECM.

4 SYSTEM EFFECTS ON ECM

Manufacturing systems are used to plan manufacturing, develop all of the subsystems, and then control the activities of manufacturing. They have a great impact on ECM. In fact, it is doubtful that ECM could be accomplished without active manufacturing system participation. At the metasystem levels, there must be an overt commitment to ECM. A company is driven by risk reduction, corporate image, and economic objectives. The leaders of the corporation must understand and accept that ECM can reduce risk, improve corporate image, and improve the economics of the firm.⁵ Without such an understanding at the top decision-making levels, ECM stands little chance of being fully implemented. A company pursuing ECM should state that ECM is part of the core business structure. This has the influence of directing the subordinate-level activities toward choices that support ECM.

In addition, the *make-to-order* and the *make-to-stock* decisions influence the amount of material that a firm must hold to support its operations. Because of external constraints on certain materials and lead times for certain components, there will be differences in what manufacturing strategy a corporation will take. However, anytime that a company must store raw materials, components, or finished product, there is an increased opportunity for waste to occur. If these materials also include hazardous materials, handling awareness also increases costs.

At the strategic level, ECM is most affected by capacity and facility planning. Facilities determine the processes available to the firm. They also determine the capacity. Process selection for a facility has the greatest influence on whether ECM can be achieved in a given facility. Any process that requires auxiliary materials to achieve an objective is subject for consideration for improvement under ECM. For example, a process that requires hydrochloric acid to clean a metal surface of rust and preservatives must be carefully examined. There are alternative processes that do not require the acid. However, once the process is selected and built into a facility, cost of replacement, space limitations, and a host of other factors may render it impossible to replace the process and therefore limit the effectiveness of subsequent ECM attempts.

Additionally, at the strategic level, choices are made on the layout of a given facility. Traditionally, these are transfer line, cell or flexible machine layouts, colony or job shop process layout, and (in the most extreme case) stationary product layout. This choice must satisfy the given production requirements that include product, quality, and process flow. Choice of the wrong layout strategy results in increased energy consumption, increased work in progress, increased internal plant transportation, and increased waste.

The second greatest influence on ECM occurs at the tactical level in process planning. Process planning is the conversion of the engineering design to definition of the detailed processes that will enable creation of the product. Process planning can be considered at two levels:

1. The *macrolevel* where decisions are made as to what process of the facility will be used and in what order
2. The *microlevel*, where all of the details of machines, individual tools, and instructions are created

Because of the influence of process planning on the individual details of creating a product, choices here directly influence waste, use of hazardous materials, and energy consumption. As the old cliché goes, “The devil is in the details!” Process planners intimate with the facility can greatly reduce waste and energy consumption by choice of energy saving and near-net-shape processes. An additional benefit for this is that the cost of production is also less.

Another tactical decision that can have an influence on ECM is production scheduling. Long-batch-run schedules tend to waste less and consume less energy. Process setups and changeovers are notoriously wasteful and costly. In addition, much of the quality control wastage occurs during the early part of new batch runs. However, the same benefits can often be achieved through planning the facility for flexibility at the strategic level.

At the lowest levels, controls have the least impact on ECM, although significant improvements are still possible. For example, early detection of a faulty process can prevent excess material and energy use. This is even more critical if the process in question is near the beginning of the product’s manufacturing.

The controls are often the best indicators of inadequate ECM decisions made at higher levels, even though they have the least impact. For example, while monitoring the production process, shop floor foremen should be aware of places where waste collects and should report this to the process planners and process engineers. Machine maintenance is often an indicator: Frequent stoppages and unplanned maintenance are an indication of an improper process and of waste. Thus, assessment activities focus on data and observations at these levels. But where problems exist, oftentimes the improvement will be in a higher level system.

This section discussed illustrations of systematic effects on ECM. The next section focuses on assessment and evaluation and improvement methods for achieving ECM.

5 ASSESSMENT

Assessment as used here is the examination and evaluation of the effectiveness of a manufacturing facility and its systems in meeting the objectives of ECM. Assessment is properly the *check* phase of PDCA. In assessment we examine the processes that are occurring and determine how well they function. The purpose of assessment is to benchmark the current state of operations against an ideal state of operations. Plants will have both exemplary operations and operations that do not meet desired standards. It is essential to recognize the exemplary operations for several purposes: Exemplary practices should be copied where feasible across the corporation. In addition, recognition of what is being performed right is important to encouraging and rewarding the plant management for making improvements.

There are three parts to assessment: data collection, evaluation, and reporting. During data collection, information is gathered and analyzed to determine how the plant is performing over an extended period, usually a year or more. The evaluation is conducted on-site and is a snapshot of where the plant is at the time of assessment. Reporting is essential in documenting the assessment procedures, the assessment recommendations, and the plant exemplary practices so that they can be communicated appropriately.

5.1 Assessment Planning

The time and effort required for assessment will vary, depending on the size of the organization being evaluated. Typically, an assessment team will consist of a team leader and three to five team members. The team leader should have familiarity with the plant and its operations. However, the team leader and its members must be independent of the plant in the corporation organization. It is most common to select members of an industrial staff from another plant to perform the assessment if no formal organization exists in the corporate structure. For small firms with only one plant, these requirements cannot be met; however, assessments can still be performed either using outside resources or by subdividing the plant and limiting the assessments to subunits. An excellent program to assist small companies can be referenced at <http://www1.eere.energy.gov/industry/bestpractices/iacs.html>.³ Often the team leader is a senior industrial engineer. The team members are normally engineers and planners for the organization.

A typical assessment will require about a week to perform provided the team is trained and has performed previous assessments. This will be far less for small plants and firms having only one plant, but the proportion of time planned will be approximately the same. The first three days of the assessment concentrates on data collection and analysis. The fourth day is the on-site plant visit. The fifth day is the on-site evaluation and report writing.

It is not uncommon and in some cases is desired that the first three days be separated from the plant visit by a time period of up to a month. This is to allow more time for analysis and to ensure that the correct data have been provided. It is desirable to have actual data, not summaries. However, the manufacturing unit will often try to provide summary only: This must be resisted. The actual data will have timing information and anomalies that are usually absent from summarized data. Where summaries have been furnished, time will be needed to make requests for the actual data and to receive it in time for the analysis to be completed before the site visit.

The first part of the assessment is analysis of the documentation of a plant. Depending on the plant, this may be performed at the plant location or at another corporate location. Off-site evaluation is preferred to on-site evaluation to reduce potential interference with plant operations, reduce assessment cost, and provide the team with the most time to analyze the data.

On-site visits are usually less productive because of the need for the plant management to perform briefings, set up work areas, and preinterview team members from the concerned staff. During the first two hours of the first day, the team leader assigns team members to analyze specific data. One team member should focus on plant processes, quality, and scheduling. Another team member should focus on plant purchases and inventory. Yet another team member should concentrate on plant emissions and hazardous substances. A team member should be assigned to studying energy, water, and solid wastes for the plant. This analysis will focus on specific manufacturing systems improvement opportunities.

At the end of the third day, the team meets for approximately two hours to discuss findings, compare notes, and establish what areas are of concern for the on-site visit. Team members should report what they have analyzed, what were the results, and any major omissions that may have occurred. The team leader may ask certain team members to concentrate on certain parts of the facility based on these findings.

In addition to accessing documents, data collection also includes visiting the processes or facilities being assessed and interviewing managers and employees. The visit is organized by arrival on the site at the beginning of the work day. A management briefing on the plant will review the layout, processes, and safety requirements for the plant. This is usually followed by a management-guided walkthrough of the plant. The management briefings and the management-guided walkthrough should be performed in approximately two hours. The team leader then assigns members areas of closer inspections and interviews in areas of interest.

At midday, the team should meet in a private location for the midday meal and should discuss any significant findings. Following this meal, the team leader may make additional assignments. The next three hours should be used to complete the site inspection and to perform any more interviews needed. In multishift operations, assessment team members may be required to attend those shifts for interviews and inspections. Often in a multishift operation, certain operations are only performed in the night shift, or the graveyard shift.

The first hour of the fifth day should be used to formulate any major findings that will be in the team report. At the end of this hour, the plant management is presented with these findings for their further comments. After hearing the plant management, the next two hours are used to discuss the formulation of recommended improvement projects. The remainder of the fifth day is used to formalize the team report. The formal report should be completed within two weeks of the site visit.

5.2 Data Collection

During the data collection phase of assessment, an attempt is made to collect all of the documents relevant to the operating state. Usually, a year's worth of consecutive data is needed. Information needed relates to operations, material storage and use, identified hazardous materials, and energy usage. This would include production data in units of production, aggregate plans, production schedules, downtime reports, material safety data sheets (MSDSs), quality defect reports (QDRs), purchase orders for materials, shipping orders, inventory turnover reports, emission reports, and energy bills. In some cases, the records will be too voluminous to copy and the data may have to be examined at the plant location for the usage of the data. Most facilities now keep electronic data rather than paper records. If not, this may be a big area for improving ECM. In addition, if previous assessment reports exist, these should be incorporated into the data analysis and used for comparative purposes.

When electronic records exist, usually the ability to perform statistical analysis is much easier. Basic statistics such as means and variances should be computed and compared to benchmarking statistics if available. In addition, trend analysis may be useful to find processes that may be approaching a critical situation. Special care should be given to data that seem unusual when taken in the context of the complete data stream. Outliers are often improper measurements; however, they may also flag conditions that require further action.

Production schedules indicate how often the production is being changed. In addition, they will indicate by the plan date how much lead time is given to changing the schedule. Short planning lead times are an indicator of poor planning and waste. Downtime reports will provide information on the reliability of a process. Planned schedules should be compared to actual performance to find discrepancies. Where these exist, they should be investigated to determine the reason. QDRs are particularly important, as these will highlight problematic areas with the production line. Most firms today use some form of statistical quality control. If this is not in use, this will flag a very important area for improvement. Excessive QDRs often highlight process planning problems.

Material purchase orders are measures of the inputs that the system is actually consuming. This needs to be compared to shipping orders to determine whether the material ordered is being converted to product. It is particularly important in the collecting and analyzing of purchase orders and shipping orders to observe waste collection volume, type, and frequency. Inventory turnover reports will indicate materials that are stagnant within the system. A high percentage of nonrotating stockages indicates poor production scheduling, poor material planning, or an inability of managers to release one-time stocks once acquired. The result is waste in time, storage, and capital to support maintenance of this inventory.

MSDSs will indicate the hazardous materials in use as well as the handling requirements. The person(s) doing the assessment should observe in the site visit where the materials are and how they are being handled. The emission reports from the facility indicate what substances are being released into the atmosphere. Because of the cost of licensing and compliance, any savings here can have a big impact on ECM.

The energy bills will indicate how energy is used and what types of loads are being billed against. Computation of power (the ratio of real power to total power) and load factors (overall percentage use of electrical equipment) should be performed and graphed. Power factors below 95% should be flagged as opportunities for improvement for reducing energy costs. Load factors below 70% indicate that equipment is not being used near its peak operating requirements. When these data are taken in their entirety, a picture will form of the state of current operations. In addition, inappropriate load factors will be highlighted that are out of place and will become opportunities for improvements. More extensive energy auditing information is located at the Energy Star website, <http://www.energystar.gov/>, or the more specific link, http://www.energystar.gov/index.cfm?c=industry.bus_industry_plant_energy_auditing.⁴

5.3 Site Visit and Inspection

The site visit begins with a plant management briefing and plant management-guided walk-through. During the management briefing, often in the question-and-answer period after the formal briefing, the assessment team should relate any significant findings that may have been noticed in the data and should ask questions regarding these. This discussion will prepare the management staff for the interviews to follow later in the day so they can gather the relevant information.

Information that was found in the documentation should be confirmed on the floor of the facility wherever possible. During the visit, carefully note the state of the operations. Is the facility adequately lighted? Is the facility operating in an orderly manner? Are the floors and the machines clean? What smells exist and from where do they originate? Is work in progress entering the areas reserved for transportation and passage? Are you receiving the same messages from the employees and the managers when you interview them and discuss their operations?

In addition, the assessment team member should walk the outside of the facility. It is often surprising what is found on the perimeter of the property of a plant. The team member might spot excess materials, defective product, liquid wastes, and other indicators of ECM defects. In the past, these were often unreported, and they hid errors. These need to be investigated and cleaned up.

5.4 Reporting and Project Formulation

The report is formulated in the following outline:

1. Title page
 - a. Name of the plant assessed
 - b. Name and contact details of who the report was prepared for
 - c. Assessment report number
 - d. Date of the report
 - e. Name and contact details of the assessment team leader
2. Executive summary
3. Assessment recommendations
 - a. Title
 - b. Observed problem
 - c. Recommendation
 - d. Estimated costs and benefits
4. Exemplary plant practices
 - a. Title
 - b. Observed practice
 - c. Estimated benefits
 - d. Contact information for more details
5. Synopsis of assessment
 - a. Plant background
 - b. General plant information
 - c. Plant leadership
 - d. Plant processes
 - e. Description of wastes
 - f. Description of energy usage
 - g. Hazardous materials in use
6. Data analysis—synopsis only (details are in appendices)
7. Site visit observations
 - a. What was observed
 - b. Who observed it
 - c. When it was observed
 - d. Why it is significant
 - e. Corrective action or plant response, if any
8. Findings and conclusions
9. Appendices as needed

The most important parts of the assessment report are the assessment recommendations and the exemplary plant practices. Each assessment recommendation (AR) should be itemized. The ARs provide the foundation for creating an improvement project. To do this, an AR should be titled in an active manner, starting with an action verb. An example is “Meter water use to reduce sewage charges.” This then should be followed by a description of anticipated benefits

and savings by performing the action. Where possible, this should cite both dollars in savings, as well as fundamental units such as kilowatt-hours per year or pounds of CO₂. This is followed by who observed the operation and what exactly was observed. The team member should take or obtain pictures to illustrate the process recommended for change. Cost of the improvement project should be estimated if possible, with a short financial analysis of payback.

A key component of assessment is two-way communication between the assessment team and the plant management. In the process, there are ample opportunities for this to occur. On the fourth day, the plant management team provides information to the team and may wish certain areas to be emphasized in the study and may make certain recommendations for improvements. The team has discussed with the management staff the findings from the documentation. On the fifth day, the team reports its significant findings to the plant management team for comments.

Exemplary plant practices—or *best practices*, as they are often titled—have a similar format. The title should indicate an action: “XYZ Plant uses ultrasonic tank cleaning.” Benefits achieved are itemized. Then the report should give a short description of the practice followed by a contact person in the plant who can provide more information.

The final report must be timely. The report should be ready within two weeks of the site visit and should be provided to the plant management team and the corporate vice president for the division. This may result in an order for a follow-up assessment to ensure that the assessment recommendations are being implemented and corrective actions were taken where needed.

6 SUMMARY

Manufacturing system evaluation for ECM is based on three overriding goals:

1. Reduce waste.
2. Reduce hazardous materials and processes.
3. Reduce energy.

These goals not only improve the responsiveness of the manufacturing unit to the environment but also result in significant cost savings and a better product when performed properly. Evaluation of manufacturing systems for environmentally conscious design begins with an assessment that highlights significant opportunities for improvement. Although the assessment focuses on actual operations, the solutions must be applied to the systems that created, support, and monitor the operations. Whereas ISO 14001 is a worthy goal to pursue and may be essential to the business practices of the firm, it is largely a paperwork exercise that does not solve common problems in manufacturing and does not necessarily lead to continuous improvement of the plant, the product, or the manufacturing systems. The evaluation process described in this chapter has been applied to several hundred different plants in the United States by a large number of auditors and assessors and has been shown to lead to real results in a timely fashion. It can also assist the ISO 14001 aspirations of the firm, if used properly.

REFERENCES

1. A. Ghisellinia and D. L. Thurston, “Decision Traps in ISO 14001 Implementation Process: Case Study Results from Illinois Certified Companies,” *J. Cleaner Production*, **13**, 763–777, 2005.
2. W. W. Olson, “Systems Thinking,” in M. A. Abraham (Ed.), *Sustainability Science and Engineering*, Elsevier, Amsterdam, 2006, p. 93.
3. U.S. Department of Energy Industrial Assessment Program, available: <http://www1.eere.energy.gov/industry/bestpractices/iacs.html>.
4. U.S. Energy Star, available: <http://www.energystar.gov/> and http://www.energystar.gov/index.cfm?c=industry.bus_industry_plant_energy_auditing.

CHAPTER 6

METAL FORMING, SHAPING, AND CASTING

Magd E. Zohdi and William E. Biles
University of Louisville
Louisville, Kentucky

1 INTRODUCTION	195	4.3 Permanent-Mold Casting	222
2 HOT-WORKING PROCESSES	195	4.4 Plaster Mold Casting	223
2.1 Classification of Hot-Working Processes	197	4.5 Investment Casting	224
2.2 Rolling	197	5 PLASTIC MOLDING PROCESSES	224
2.3 Forging	199	5.1 Injection Molding	224
2.4 Extrusion	201	5.2 Coinjection Molding	225
2.5 Drawing	203	5.3 Rotomolding	225
2.6 Spinning	206	5.4 Expandable-Bead Molding	225
2.7 Pipe Welding	206	5.5 Extruding	225
2.8 Piercing	207	5.6 Blow Molding	225
3 COLD-WORKING PROCESSES	207	5.7 Thermoforming	226
3.1 Classification of Cold-Working Operations	208	5.8 Reinforced-Plastic Molding	226
3.2 Squeezing Processes	209	5.9 Forged-Plastic Parts	226
3.3 Bending	210	6 POWDER METALLURGY	226
3.4 Shearing	213	6.1 Properties of PM Products	227
3.5 Drawing	215	7 SURFACE TREATMENT	227
4 METAL CASTING AND MOLDING PROCESSES	218	7.1 Cleaning	227
4.1 Sand Casting	218	7.2 Coatings	230
4.2 Centrifugal Casting	220	7.3 Chemical Conversions	232
		BIBLIOGRAPHY	233

1 INTRODUCTION

Metal-forming processes use a remarkable property of metals—their ability to flow plastically in the solid state without concurrent deterioration of properties. Moreover, by simply moving the metal to the desired shape, there is little or no waste. Figure 1 shows some of the metal-forming processes. Metal-forming processes are classified into two categories: hot-working processes and cold-working processes.

2 HOT-WORKING PROCESSES

Hot working is defined as the plastic deformation of metals above their recrystallization temperature. Here it is important to note that the crystallization temperature varies greatly with different materials. Lead and tin are hot worked at room temperature, while steels require

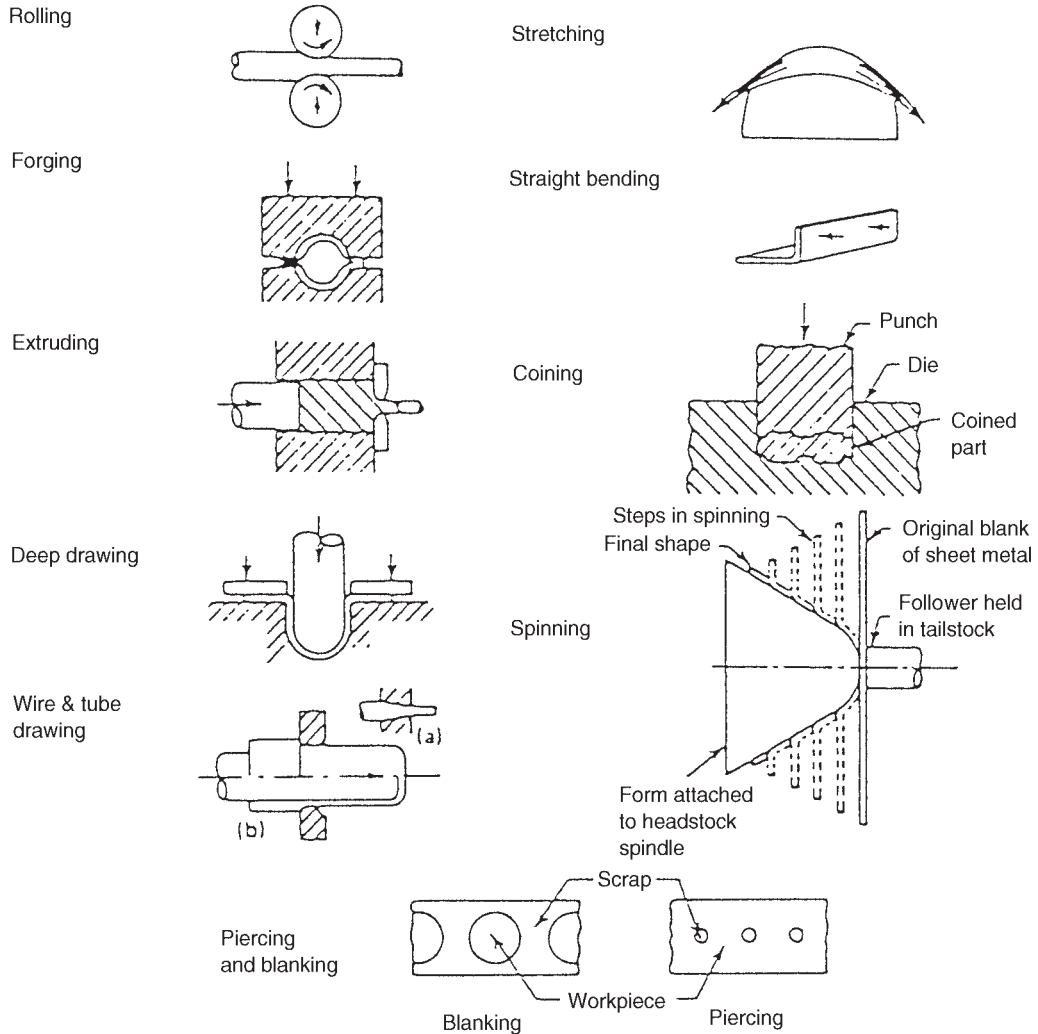


Figure 1 Metal-forming processes.

temperatures of 2000°F (1100°C). Hot working does not necessarily imply high absolute temperatures.

Hot working can produce the following improvements:

1. Production of randomly oriented, spherical-shaped grain structure, which results in a net increase not only in strength but also in ductility and toughness.
2. The reorientation of inclusions or impurity material in metal. The impurity material often distorts and flows along with the metal.

This material, however, does not recrystallize with the base metal and often produces a fiber structure. Such a structure clearly has directional properties, being stronger in one direction

than in another. Moreover, an impurity originally oriented so as to aid crack movement through the metal is often reoriented into a “crack-arrestor” configuration perpendicular to crack propagation.

2.1 Classification of Hot-Working Processes

The most obvious reason for the popularity of hot working is that it provides an attractive means of forming a desired shape. Some of the hot-working processes that are of major importance in modern manufacturing are

1. Rolling
2. Forging
3. Extrusion and upsetting
4. Drawing
5. Spinning
6. Pipe welding
7. Piercing

2.2 Rolling

Hot rolling (Fig. 2) consists of passing heated metal between two rolls that revolve in opposite directions, the space between the rolls being somewhat less than the thickness of the entering metal. Many finished parts, such as hot-rolled structural shapes, are completed entirely by hot rolling. More often, however, hot-rolled products, such as sheets, plates, bars, and strips, serve as input material for other processes, such as cold forming or machining.

In hot rolling, as in all hot working, it is very important that the metal be heated uniformly throughout to the proper temperature, a procedure known as *soaking*. If the temperature is not uniform, the subsequent deformation will also be nonuniform, the hotter exterior flowing in preference to the cooler and, therefore, stronger, interior. Cracking, tearing, and associated problems may result.

Isothermal Rolling

The ordinary rolling of some high-strength metals, such as titanium and stainless steels, particularly in thicknesses below about 0.150 in. (3.8 mm), is difficult because the heat in the sheet is transferred rapidly to the cold and much more massive rolls. This has been overcome by isothermal rolling. Localized heating is accomplished in the area of deformation by the passage of a large electrical current between the rolls, through the sheet. Reductions up to 90% per roll have been achieved. The process usually is restricted to widths below 2 in. (50 mm).

The rolling strip contact length is given by

$$L \simeq \sqrt{R(h_0 - h)}$$

where R = roll radius
 h_0 = original strip thickness
 h = reduced thickness

The roll-force F is calculated by

$$F = LwY_{\text{avg}} \quad (1)$$

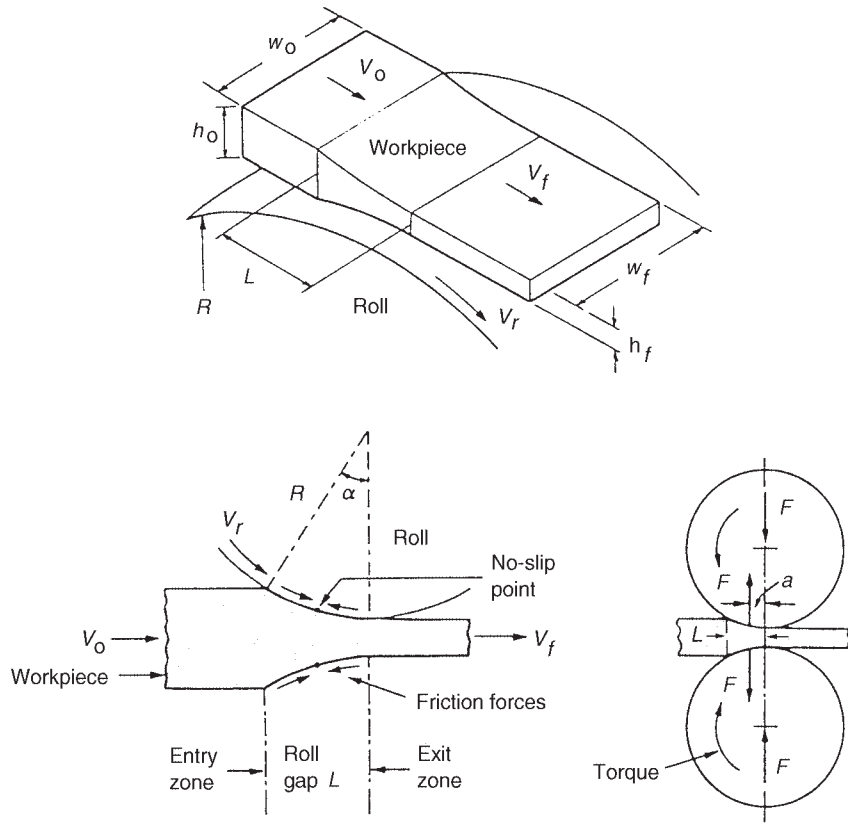


Figure 2 Hot rolling.

where w = width
 Y_{avg} = average true stress

Figure 3 gives the true stress for different material at the true stress f . The true stress f is given by

$$\epsilon = \ln \left(\frac{h_0}{h} \right)$$

$$\text{Power/Roll} = \frac{2\pi FLN}{60,000} \quad \text{kW} \quad (2)$$

where F = newtons
 L = meters
 N = rev per min

or

$$\text{Power} = \frac{2\pi FLN}{33,000} \quad \text{hp} \quad (3)$$

where F = lb
 L = ft

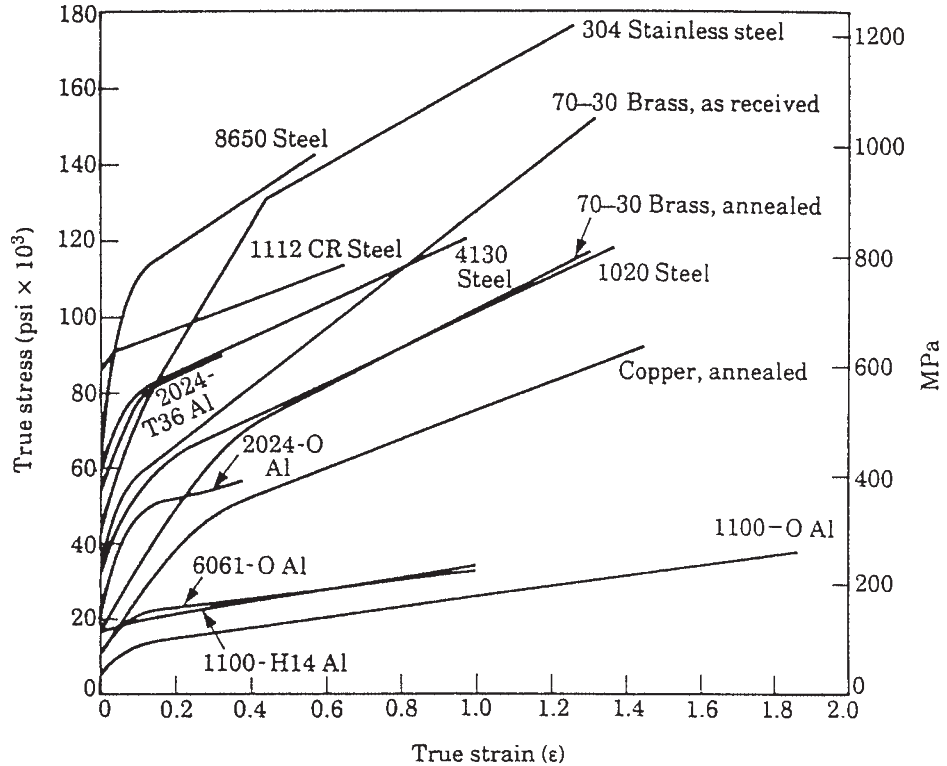


Figure 3 True stress–true strain curves.

2.3 Forging

Forging is the plastic working of metal by means of localized compressive forces exerted by manual or power hammers, presses, or special forging machines.

Various types of forging have been developed to provide great flexibility, making it economically possible to forge a single piece or to mass produce thousands of identical parts. The metal may be

1. Drawn out, increasing its length and decreasing its cross section
2. Upset, increasing the cross section and decreasing the length, or
3. Squeezed in closed impression dies to produce multidirectional flow

The state of stress in the work is primarily uniaxial or multiaxial compression.

The common forging processes are

1. Open-die hammer
2. Impression-die drop forging
3. Press forging
4. Upset forging
5. Roll forging
6. Swaging

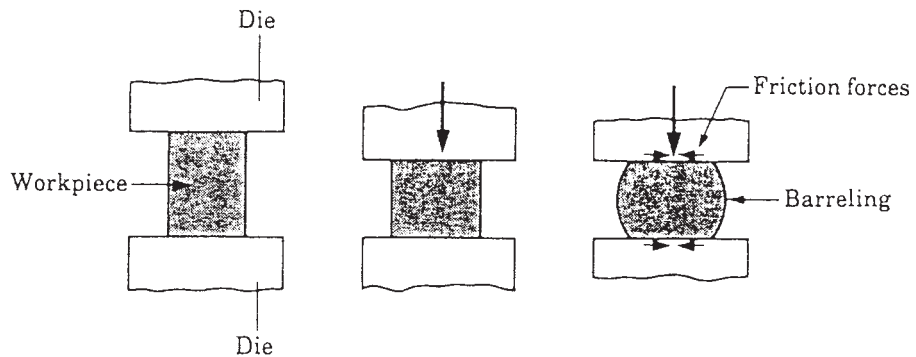


Figure 4 Open-die hammer forging.

Open-Die Hammer Forging

Open-die forging, (Fig. 4) does not confine the flow of metal, the hammer and anvil often being completely flat. The desired shape is obtained by manipulating the workpiece between blows. Specially shaped tools or a slightly shaped die between the workpiece and the hammer or anvil are used to aid in shaping sections (round, concave, or convex), making holes, or performing cutoff operations.

The force F required for an open-die forging operation on a solid cylindrical piece can be calculated by

$$F = Y_f \pi r^2 \left(1 + \frac{2\mu r}{3h} \right) \quad (4)$$

where Y_f = flow stress at the specific ϵ [$\epsilon = \ln(h_0/h)$]
 μZ = coefficient of friction
 r and h = radius and height of workpiece

Impression-Die Drop Forging

In impression-die or closed-die drop forging (Fig. 5), the heated metal is placed in the lower cavity of the die and struck one or more blows with the upper die. This hammering causes the metal to flow so as to fill the die cavity. Excess metal is squeezed out between the die faces along the periphery of the cavity to form a flash. When forging is completed, the flash is trimmed off by means of a trimming die.

The forging force F required for impression-die forging can be estimated by

$$F = KY_f A \quad (5)$$

where K = multiplying factor (4–12) depending on the complexity of the shape
 Y_f = flow stress at forging temperature
 A = projected area, including flash

Press Forging

Press forging employs a slow-squeezing action that penetrates throughout the metal and produces a uniform metal flow. In hammer or impact forging, metal flow is a response to the energy in the hammer–workpiece collision. If all the energy can be dissipated through flow of the surface layers of metal and absorption by the press foundation, the interior regions of the workpiece can go undeformed. Therefore, when the forging of large sections is required, press forging must be employed.

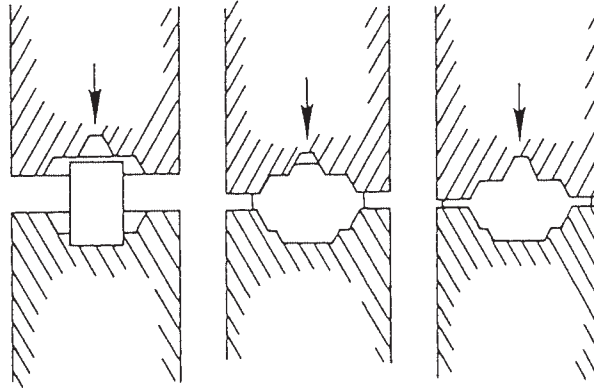


Figure 5 Impression-die drop forging.

Upset Forging

Upset forging involves increasing the diameter of the end or central portion of a bar of metal by compressing its length. Upset-forging machines are used to forge heads on bolts and other fasteners, valves, couplings, and many other small components.

Roll Forging

Roll forging, in which round or flat bar stock is reduced in thickness and increased in length, is used to produce such components as axles, tapered levers, and leaf springs.

Swaging

Swaging involves hammering or forcing a tube or rod into a confining die to reduce its diameter, the die often playing the role of the hammer. Repeated blows cause the metal to flow inward and take the internal form of the die.

2.4 Extrusion

In the extrusion process (Fig. 6), metal is compressively forced to flow through a suitably shaped die to form a product with reduced cross section. Although it may be performed either hot or cold, hot extrusion is employed for many metals to reduce the forces required, to eliminate cold-working effects, and to reduce directional properties. The stress state within the material is triaxial compression.

Lead, copper, aluminum, and magnesium, and alloys of these metals, are commonly extruded, taking advantage of the relatively low yield strengths and extrusion temperatures. Steel is more difficult to extrude. Yield strengths are high and the metal has a tendency to weld to the walls of the die and confining chamber under the conditions of high temperature and pressures. With the development and use of phosphate-based and molten glass lubricants, substantial quantities of hot steel extrusions are now produced. These lubricants adhere to the billet and prevent metal-to-metal contact throughout the process.

Almost any cross-section shape can be extruded from the nonferrous metals. Hollow shapes can be extruded by several methods. For tubular products, the stationary or moving mandrel process is often employed. For more complex internal cavities, a spider mandrel or torpedo die is used. Obviously, the cost for hollow extrusions is considerably greater than for

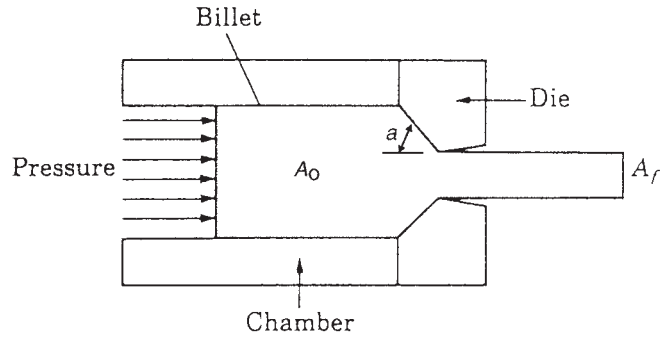


Figure 6 Extrusion process.

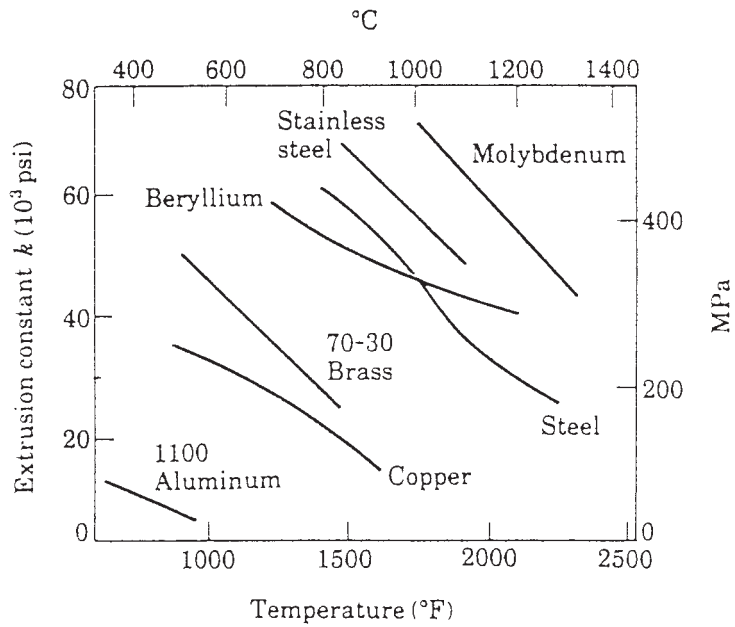


Figure 7 Extrusion constant k .

solid ones, but a wide variety of shapes can be produced that cannot be made by any other process.

The extrusion force F can be estimated from the formula

$$F = A_0 k \ln \left(\frac{A_0}{A_f} \right) \tag{6}$$

where k = extrusion constant that depends on material and temperature (see Fig. 7)

A_0 = billet area

A_f = finished extruded area

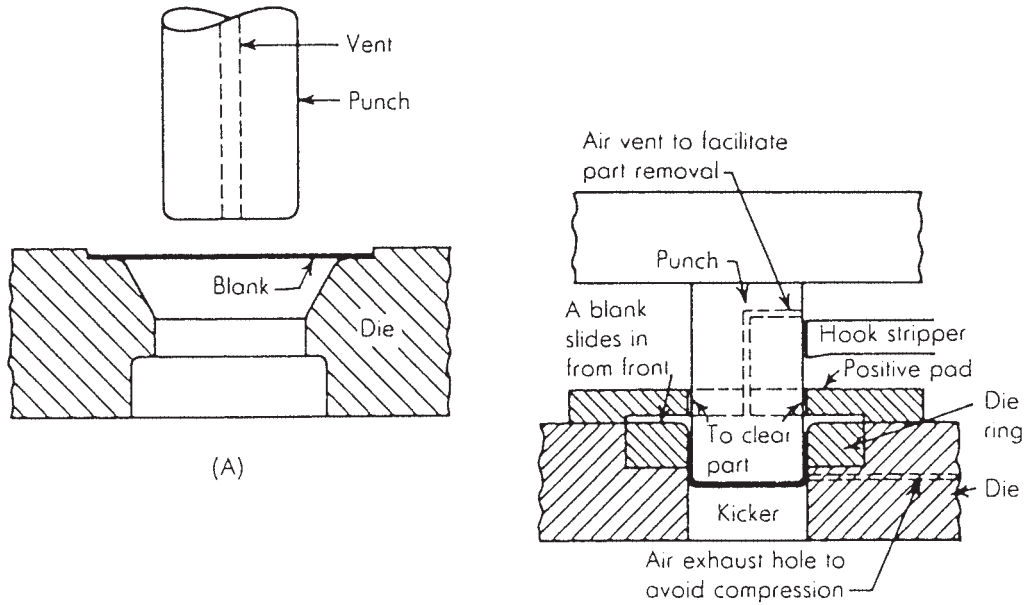


Figure 8 Drawing process.

2.5 Drawing

Drawing (Fig. 8) is a process for forming sheet metal between an edge-opposing punch and a die (draw ring) to produce a cup, cone, box, or shell-like part. The work metal is bent over and wrapped around the punch nose. At the same time, the outer portions of the blank move rapidly toward the center of the blank until they flow over the die radius as the blank is drawn into the die cavity by the punch. The radial movement of the metal increases the blank thickness as the metal moves toward the die radius; as the metal flows over the die radius, this thickness decreases because of the tension in the shell wall between the punch nose and the die radius and (in some instances) because of the clearance between the punch and the die.

The force (load) required for drawing a round cup is expressed by the following empirical equation:

$$L = \pi dtS \left(\frac{D}{d} - k \right) \quad (7)$$

where L = press load, lb

d = cup diameter, in.

D = blank diameter, in.

t = work-metal thickness, in.

S = tensile strength, lb/in.²

k = constant that takes into account frictional and bending forces, usually 0.6–0.7

The force (load) required for drawing a rectangular cup can be calculated from the following equation:

$$L = tS(2\pi Rk_A + lk_B) \quad (8)$$

where L = press load, lb
 t = work-metal thickness, in.
 S = tensile strength, lb/in.²
 R = corner radius of the cup, in.
 l = sum of the lengths of straight sections of the sides, in.
 k_A and k_B = constants

Values for k_A range from 0.5 (for a shallow cup) to 2.0 (for a cup of depth five to six times the corner radius). Values for k_B range from 0.2 (for easy draw radius, ample clearance, and no blank-holding force) and 0.3 (for similar free flow and normal blank-holding force of about $L/3$) to a maximum of 1.0 (for metal clamped too tightly to flow).

Figure 9 can be used as a general guide for computing maximum drawing load for a round shell. These relations are based on a free draw with sufficient clearance so that there is no ironing, using a maximum reduction of 50%. The nomograph gives the load required to fracture the cup (1 ton = 8.9 kN).

Blank Diameters

The following equations may be used to calculate the blank size for cylindrical shells of relatively thin metal. The ratio of the shell diameter to the corner radius (d/r) can affect the blank diameter and should be taken into consideration. When d/r is 20 or more,

$$D = \sqrt{d^2 + 4dh} \quad (9)$$

When d/r is between 15 and 20,

$$D = \sqrt{d^2 + 4dh - 0.5r} \quad (10)$$

When d/r is between 10 and 15,

$$D = \sqrt{d^2 + 4dh - r} \quad (11)$$

When d/r is below 10,

$$D = \sqrt{(d - 2r)^2 + 4d(h - r) + 2\pi r(d - 0.7r)} \quad (12)$$

where D = blank diameter
 d = shell diameter
 h = shell height
 r = corner radius

The above equations are based on the assumption that the surface area of the blank is equal to the surface area of the finished shell.

In cases where the shell wall is to be ironed thinner than the shell bottom, the volume of metal in the blank must equal the volume of the metal in the finished shell. Where the wall thickness reduction is considerable, as in brass shell cases, the final blank size is developed by trial. A tentative blank size for an ironed shell can be obtained from the equation

$$D = \sqrt{d^2 + 4dh \frac{t}{T}} \quad (13)$$

where t = wall thickness
 T = bottom thickness

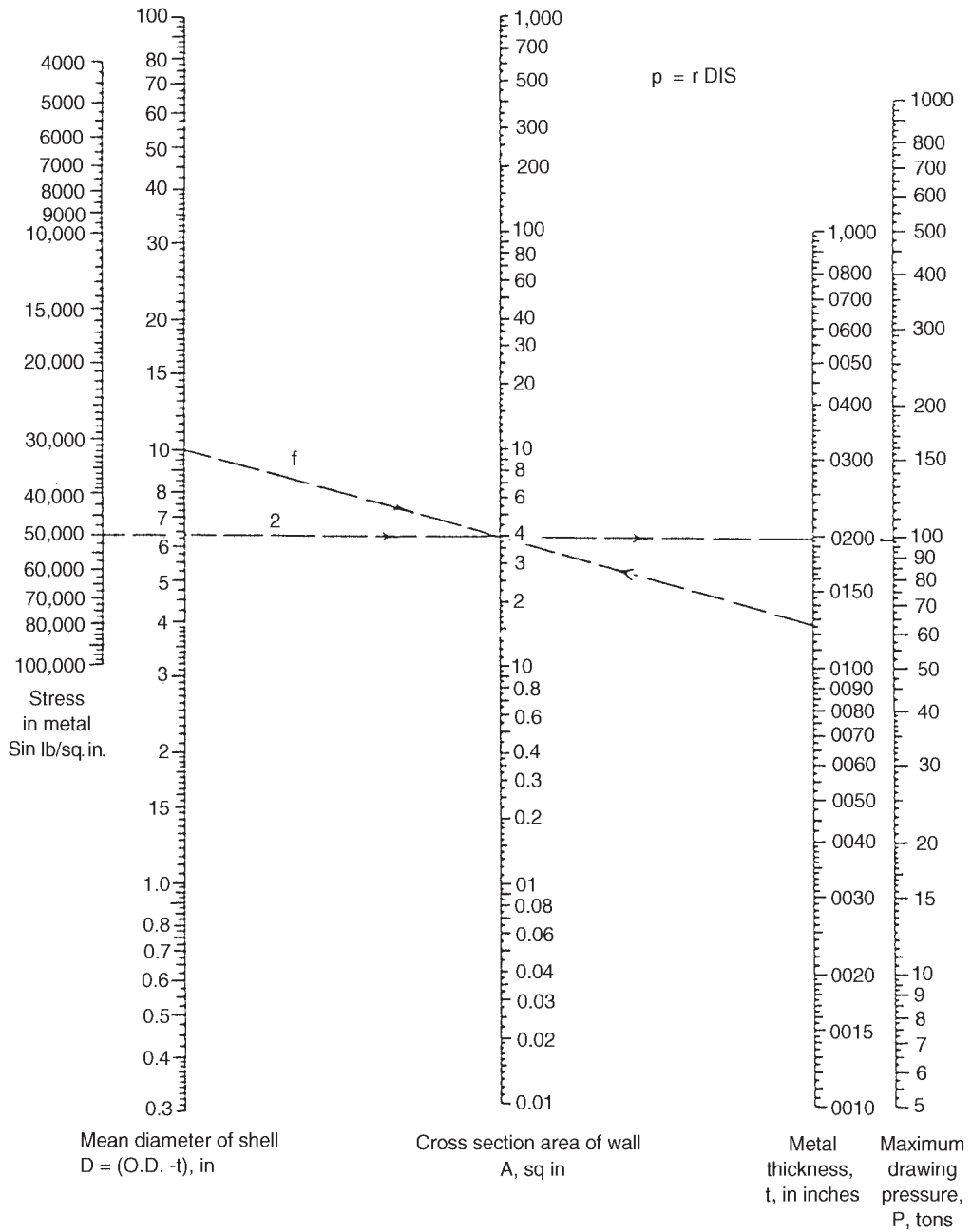


Figure 9 Nomograph for estimating drawing pressures.

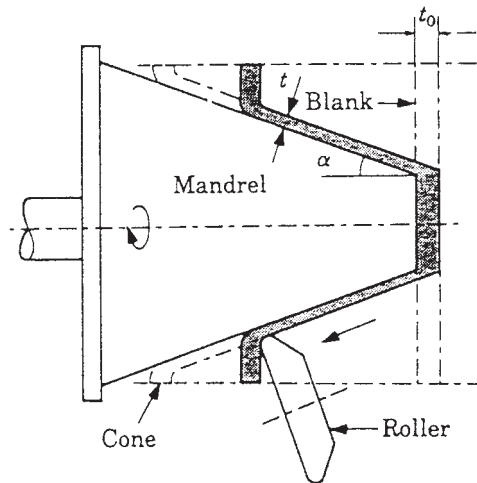


Figure 10 Setup and dimensional relations for one-operation power spinning of a cone.

2.6 Spinning

Spinning is a method of forming sheet metal or tubing into seamless hollow cylinders, cones, hemispheres, or other circular shapes by a combination of rotation and force. On the basis of techniques used, applications, and results obtainable, the method may be divided into two categories: *manual spinning* (with or without mechanical assistance to increase the force) and *power spinning*.

Manual spinning entails no appreciable thinning of metal. The operation ordinarily done in a lathe consists of pressing a tool against a circular metal blank that is rotated by the headstock.

Power spinning is also known as *shear spinning* because in this method metal is intentionally thinned, by shear forces. In power spinning, forces as great as 400 tons are used.

The application of shear spinning to conical shapes is shown schematically in Fig. 10. The metal deformation is such that forming is in accordance with the sine law, which states that the wall thickness of the starting blank and that of the finished workpiece are related as

$$t_2 = t_1(\sin \alpha) \quad (14)$$

where t_1 = thickness of the starting blank
 t_2 = thickness of the spun workpiece
 αZ = one-half the apex angle of the cone

Tube Spinning

Tube spinning is a rotary-point method of extruding metal, much like cone spinning, except that the sine law does not apply. Because the half-angle of a cylinder is zero, tube spinning follows a purely volumetric rule, depending on the practical limits of deformation that the metal can stand without intermediate annealing.

2.7 Pipe Welding

Large quantities of small-diameter steel pipe are produced by two processes that involve hot forming of metal strip and welding of its edges through utilization of the heat contained in the

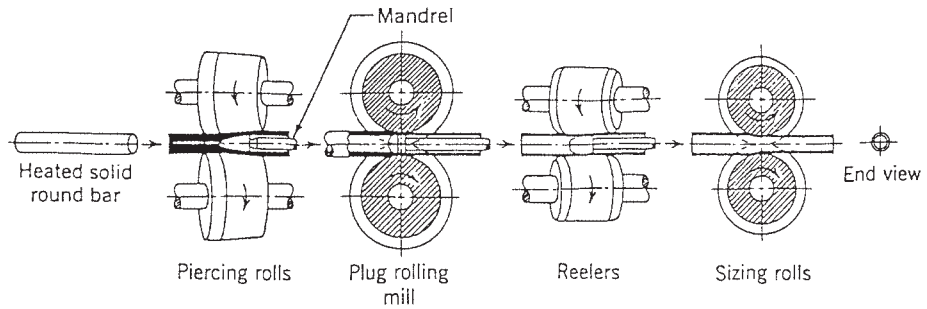


Figure 11 Principal steps in the manufacture of seamless tubing.

metal. Both of these processes, *butt welding* and *lap welding* of pipe, utilize steel in the form of skelp—long and narrow strips of the desired thickness. Because the skelp has been previously hot rolled and the welding process produces further compressive working and recrystallization, pipe welding by these processes is uniform in quality.

In the butt-welded pipe process, the skelp is unwound from a continuous coil and is heated to forging temperatures as it passes through a furnace. Upon leaving the furnace, it is pulled through forming rolls that shape it into a cylinder. The pressure exerted between the edges of the skelp as it passes through the rolls is sufficient to upset the metal and weld the edges together. Additional sets of rollers size and shape the pipe. Normal pipe diameters range from $1/8$ to 3 in. (3–75 mm).

The lap-welding process for making pipe differs from butt welding in that the skelp has beveled edges and a mandrel is used in conjunction with a set of rollers to make the weld. The process is used primarily for larger sizes of pipe, from about 2 to 14 in. (50–400 mm) in diameter.

2.8 Piercing

Thick-walled and seamless tubing is made by the piercing process. A heated, round billet, with its leading end center punched, is pushed longitudinally between two large, convex-tapered rolls that revolve in the same direction, their axes being inclined at opposite angles of about 6° from the axis of the billet. The clearance between the rolls is somewhat less than the diameter of the billet. As the billet is caught by the rolls and rotated, their inclination causes the billet to be drawn forward into them. The reduced clearance between the rolls forces the rotating billet to deform into an elliptical shape. To rotate with an elliptical cross section, the metal must undergo shear about the major axis, which causes a crack to open. As the crack opens, the billet is forced over a pointed mandrel that enlarges and shapes the opening, forming a seamless tube (Fig. 11).

This procedure applies to seamless tubes up to 6 in. (150 mm) in diameter. Larger tubes up to 14 in. (355 mm) in diameter are given a second operation on piercing rolls. To produce sizes up to 24 in. (610 mm) in diameter, reheated, double-pierced tubes are processed on a rotary rolling mill and are finally completed by reelers and sizing rolls, as described in the single-piercing process.

3 COLD-WORKING PROCESSES

Cold working is the plastic deformation of metals below the recrystallization temperature. In most cases of manufacturing, such cold forming is done at room temperature. In some cases,

however, the working may be done at elevated temperatures that will provide increased ductility and reduced strength but will be below the recrystallization temperature.

When compared to hot working, cold-working processes have certain distinct advantages:

1. No heating required.
2. Better surface finish obtained.
3. Superior dimension control.
4. Better reproducibility and interchangeability of parts.
5. Improved strength properties.
6. Directional properties can be imparted.
7. Contamination problems minimized.

Some disadvantages associated with cold-working processes include:

1. Higher forces required for deformation.
2. Heavier and more powerful equipment required.
3. Less ductility available.
4. Metal surfaces must be clean and scale free.
5. Strain hardening occurs (may require intermediate anneals).
6. Imparted directional properties may be detrimental.
7. May produce undesirable residual stresses.

3.1 Classification of Cold-Working Operations

The major cold-working operations can be classified basically under the headings of squeezing, bending, shearing, and drawing, as follows:

Squeezing	Bending	Shearing	Drawing
1. Rolling	1. Angle	1. Shearing slitting	1. Bar and tube drawing
2. Swaging	2. Roll	2. Blanking	2. Wire drawing
3. Cold forging	3. Roll forming	3. Piercing lancing perforating	3. Spinning
4. Sizing	4. Drawing	4. Notching nibbling	4. Embossing
5. Extrusion	5. Seaming	5. Shaving	5. Stretch forming
6. Riveting	6. Flanging	6. Trimming	6. Shell drawing
7. Staking	7. Straightening	7. Cutoff	7. Ironing
8. Coining		8. Dinking	8. High-energy-rate forming
9. Peening			
10. Burnishing			
11. Die hobbing			
12. Thread rolling			

3.2 Squeezing Processes

Most of the cold-working squeezing processes have identical hot-working counterparts or are extensions of them. The primary reasons for deforming cold rather than hot are to obtain better dimensional accuracy and surface finish. In many cases, the equipment is basically the same, except that it must be more powerful.

Cold Rolling

Cold rolling accounts by far for the greatest tonnage of cold-worked products. Sheets, strip, bars, and rods are cold rolled to obtain products that have smooth surfaces and accurate dimensions.

Swaging

Swaging basically is a process for reducing the diameter, tapering, or pointing round bars or tubes by external hammering. A useful extension of the process involves the formation of internal cavities. A shaped mandrel is inserted inside a tube and the tube is then collapsed around it by swaging (Fig. 12).

Cold Forging

Extremely large quantities of products are made by cold forging in which the metal is squeezed into a die cavity that imparts the desired shape. Cold heading is used for making enlarged sections on the ends of rod or wire, such as the heads on bolts, nails, rivets, and other fasteners.

Sizing

Sizing involves squeezing areas of forgings or ductile castings to a desired thickness. It is used principally on basses and flats, with only enough deformation to bring the region to a desired dimension.

Extrusion

This process is often called *impact extrusion* and was first used only with the low-strength ductile metals, such as lead, tin, and aluminum, for producing such items as collapsible tubes for toothpaste, medications, and so forth; small “cans” such as are used for shielding in electronics and electrical apparatus; and larger cans for food and beverages. In recent years, cold extrusion has been used for forming mild steel parts, often being combined with cold heading.

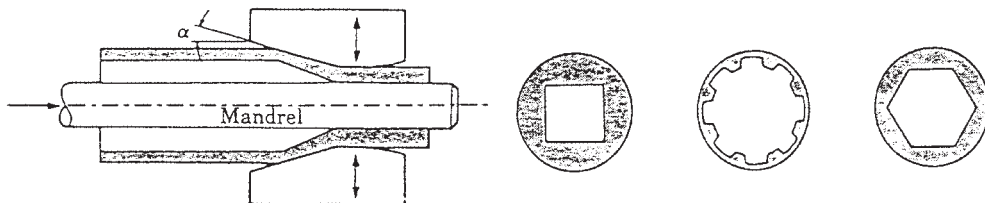


Figure 12 Cross sections of tubes produced by swaging on shaped mandrels. Rifling (spiral grooves) in small gun barrels can be made by this process.

Another type of cold extrusion, known as *hydrostatic extrusion*, used high fluid pressure to extrude a billet through a die, either into atmospheric pressure or into a lower pressure chamber. The pressure-to-pressure process makes possible the extrusion of relatively brittle materials, such as molybdenum, beryllium, and tungsten. Billet-chamber friction is eliminated, billet-die lubrication is enhanced by the pressure, and the surrounding pressurized atmosphere suppresses crack initiation and growth.

Riveting

In riveting, a head is formed on the shank end of a fastener to provide a permanent method of joining sheets or plates of metal together. Although riveting usually is done hot in structural work, in manufacturing it almost always is done cold.

Staking

Staking is a commonly used cold-working method for permanently fastening two parts together where one protrudes through a hole in the other. A shaped punch is driven into one of the pieces, deforming the metal sufficiently to squeeze it outward.

Coining

Coining involves cold working by means of positive-displacement punch while the metal is completely confined within a set of dies.

Peening

Peening involves striking the surface repeated blows by impelled shot or a round-nose tool. The highly localized blows deform and tend to stretch the metal surface. Because the surface deformation is resisted by the metal underneath, the result is a surface layer under residual compression. This condition is highly favorable to resist cracking under fatigue conditions, such as repeated bending, because the compressive stresses are subtractive from the applied tensile loads. For this reason, shafting, crankshafts, gear teeth, and other cyclic-loaded components are frequently peened.

Burnishing

Burnishing involves rubbing a smooth, hard object under considerable pressure over the minute surface protrusions that are formed on a metal surface during machining or shearing, thereby reducing their depth and sharpness through plastic flow.

Hobbing

Hobbing is a cold-working process that is used to form cavities in various types of dies, such as those used for molding plastics. A male hob is made with the contour of the part that ultimately will be formed by the die. After the hob is hardened, it is slowly pressed into an annealed die block by means of hydraulic press until the desired impression is produced.

Thread Rolling

Threads can be rolled in any material sufficiently plastic to withstand the forces of cold working without disintegration. Threads can be rolled by flat or roller dies.

3.3 Bending

Bending is the uniform straining of material, usually flat sheet or strip metal, around a straight axis that lies in the neutral plane and normal to the lengthwise direction of the sheet or strip.

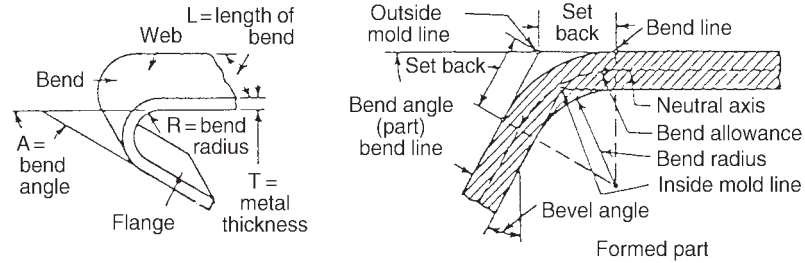


Figure 13 Bend terms.

Metal flow takes place within the plastic range of the metal, so that the bend retains a permanent set after removal of the applied stress. The inner surface of the bend is in compression; the outer surface is in tension.

Terms used in bending are defined and illustrated in Fig. 13. The neutral axis is the plane area in bent metal where all strains are zero.

Bend Allowances

Since bent metal is longer after bending, its increased length, generally of concern to the product designer, may also have to be considered by the die designer if the length tolerance of the bent part is critical. The length of bent metal may be calculated from the equation

$$B = \frac{A}{360} \times 2\pi(R_i + Kt) \quad (15)$$

where B = bend allowance, in. (mm) (along neutral axis)

A = bend angle, deg

R_i = inside radius of bend, in. (mm)

t = metal thickness, in. (mm)

$K = 0.33$ when R_i is less than $2t$ and is 0.50 when R_i is more than $2t$

Bending Methods

Two bending methods are commonly made use of in press tools. Metal sheet or strip, supported by a V block (Fig. 14), is forced by a wedge-shaped punch into the block. Edge bending (Fig. 14) is cantilever loading of a beam. The bending punch (1) forces the metal against the supporting die (2).

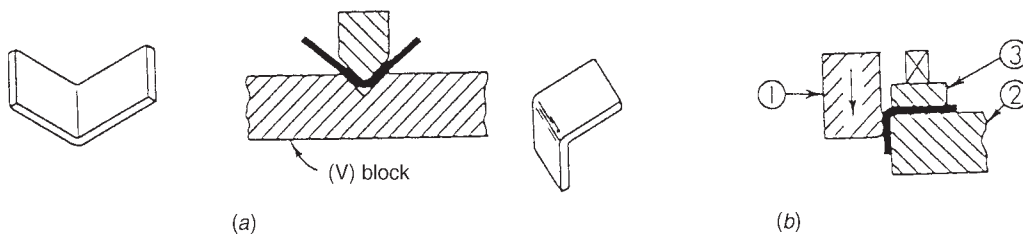


Figure 14 Bending methods (a) V bending and (b) edge bending.

Table 1 Ultimate Strength

Metal	ton/in. ²
Aluminum and alloys	6.5–38.0
Brass	19.0–38.0
Bronze	31.5–47.0
Copper	16.0–25.0
Steel	22.0–40.0
Tin	1.1–1.4
Zinc	9.7–13.5

Bending Force

The force required for V bending is

$$P = \frac{KLS^2}{W} \quad (16)$$

where P = bending force, tons (for metric usage, multiply number of tons by 8.896 to obtain kilonewtons)

K = die opening factor: 1.20 for a die opening of 16 times metal thickness, 1.33 for an opening of eight times metal thickness

L = length of part, in.

S = ultimate tensile strength, tons/in.²

W = width of V or U die, in.

t = metal thickness, in.

For U bending (channel bending), pressures will be approximately twice those required. For U bending, edge bending is required about one-half those needed for V bending. Table 1 gives the ultimate strength = S for various materials.

Several factors must be considered when designing parts that are to be made by bending. Of primary importance is the minimum radius that can be bent successfully without metal cracking. This, of course, is related to the ductility of the metal.

Angle Bending

Angle bends up to 150° in the sheet metal under about 1/16 in. (1.5 mm) in thickness may be made in a bar folder. Heavier sheet metal and more complex bends in thinner sheets are made on a press brake.

Roll Bending

Plates, heavy sheets, and rolled shapes can be bent to a desired curvature on forming rolls. These usually have three rolls in the form of a pyramid, with the two lower rolls being driven and the upper roll adjustable to control the degree of curvature. Supports can be swung clear to permit removal of a closed shape from the rolls. Bending rolls are available in a wide range of sizes, some being capable of bending plate up to 6 in. (150 mm) thick.

Cold Roll Forming

This process involves the progressive bending of metal strip as it passes through a series of forming rolls. A wide variety of moldings, channeling, and other shapes can be formed on machines that produce up to 10,000 ft (3000 m) of product per day.

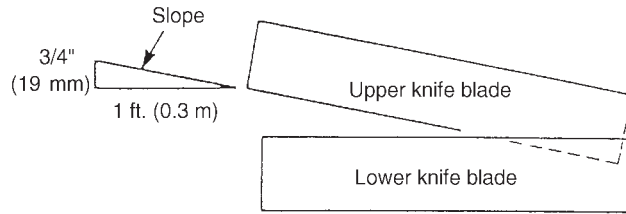


Figure 15 Rake is the angular slope formed by the cutting edges of the upper and lower knives.

Seaming

Seaming is used to join ends of sheet metal to form containers such as cans, pails, and drums. The seams are formed by a series of small rollers on seaming machines that range from small hand-operated types to large automatic units capable of producing hundreds of seams per minute in the mass production of cans.

Flanging

Flanges can be rolled on sheet metal in essentially the same manner as seaming is done. In many cases, however, the forming of flanges and seams involves drawing since localized bending occurs on a curved axis.

Straightening

Straightening or flattening has as its objective the opposite of bending and often is done before other cold-forming operations to ensure that flat or straight material is available. Two different techniques are quite common. *Roll straightening* or *roller leveling* involves a series of reverse bends. The rod, sheet, or wire is passed through a series of rolls having decreased offsets from a straight line. These bend the metal back and forth in all directions, stressing it slightly beyond its previous elastic limit and thereby removing all previous permanent set.

Sheet may also be straightened by a process called *stretcher leveling*. The sheets are grabbed mechanically at each end and stretched slightly beyond the elastic limit to remove previous stresses and thus produce the desired flatness.

3.4 Shearing

Shearing is the mechanical cutting of materials in sheet or plate form without the formation of chips or use of burning or melting. When the two cutting blades are straight, the process is called *shearing*. Other processes, in which the shearing blades are in the form of curved edges or punches and dies, are called by other names, such as *blanking*, *piercing*, *notching*, *shaving*, and *trimming*. These all are basically shearing operations, however.

The required shear force can be calculated as

$$F = \left(\frac{S \times P \times t^2 \times 12}{R} \right) \left(1 - \frac{P}{2} \right) \quad (17)$$

where F = shear force, lb

S = shear strength (stress), psi

P = penetration of knife into material, %

t = thickness of material, in.

R = rake of the knife blade, in./ft (Fig. 13)

Table 2 Values of Percent Penetration and Shear Strength for Various Materials

Material	Percent Penetration	Shear Strength, psi (MPa)
Lead alloys	50	3500 (24.1)–6000 (41.3)
Tin alloys	40	5000 (34.5)–10,000 (69)
Aluminum alloys	60	8000 (55.2)–45,000 (310)
Titanium alloys	10	60,000 (413)–70,000 (482)
Zinc	50	14,000 (96.5)
Cold worked	25	19,000 (131)
Magnesium alloys	50	17,000 (117)–30,000 (207)
Copper	55	22,000 (151.7)
Cold worked	30	28,000 (193)
Brass	50	32,000 (220.6)
Cold worked	30	52,000 (358.5)
Tobin bronze	25	36,000 (248.2)
Cold worked	17	42,000 (289.6)
Steel, 0.10C	50	35,000 (241.3)
Cold worked	38	43,000 (296.5)
Steel, 0.40C	27	62,000 (427.5)
Cold worked	17	78,000 (537.8)
Steel, 0.80C	15	97,000 (668.8)
Cold worked	5	127,000 (875.6)
Steel, 1.00C	10	115,000 (792.9)
Cold worked	2	150,000 (1034.2)
Silicon steel	30	65,000 (448.2)
Stainless steel	30	57,000 (363)–128,000 (882)
Nickel	55	35,000 (241.3)

For SI units, the force is multiplied by 4.448 to obtain newtons (N). Table 2 gives the values of P and S for various materials.

Blanking

A blank is a shape cut from flat or preformed stock. Ordinarily, a blank serves as a starting workpiece for a formed part; less often, it is a desired end product.

Calculation of the forces and the work involved in blanking gives average figures that are applicable only when (1) the correct shear strength for the material is used, and (2) the die is sharp and the punch is in good condition, has correct clearance, and is functioning properly.

The total load on the press, or the press capacity required to do a particular job, is the sum of the cutting force and other forces acting at the same time, such as the blank-holding force exerted by a die cushion.

Cutting Force: Square-End Punches and Dies

When punch and die surfaces are flat and at right angles to the motion of the punch, the cutting force can be found by multiplying the area of the cut section by the shear strength of the work material:

$$L = Stl \quad (18)$$

where L = load on the press, lb (cutting force)
 S = shear strength of the stock, psi
 t = stock thickness, in.
 l = length or perimeter of cut, in.

Piercing

Piercing is a shearing operation wherein the shearing blades take the form of closed, curved lines on the edges of a punch and die. Piercing is basically the same as blanking except that the piece punched out is the scrap and the remainder of the strip becomes the desired workpiece.

Lancing

Lancing is a piercing operation that may take the form of a slit in the metal or an actual hole. The purpose of lancing is to permit adjacent metal to flow more readily in subsequent forming operations.

Perforating

Perforating consists of piercing a large number of closely spaced holes.

Notching

Notching is essentially the same as piercing except that the edge of the sheet of metal forms a portion of the periphery of the piece that is punched out. It is used to form notches of any desired shape along the edge of a sheet.

Nibbling

Nibbling is a variation of notching in which a special machine makes a series of overlapping notches, each farther into the sheet of metal.

Shaving

Shaving is a finished operation in which a very small amount of metal is sheared away around the edge of a blanked part. Its primary use is to obtain greater dimensional accuracy, but it also may be used to obtain a square or smoother edge.

Trimming

Trimming is used to remove the excess metal that remains after a drawing, forging, or casting operation. It is essentially the same as blanking.

Cutoff

A cutoff operation is one in which a stamping is removed from a strip of stock by means of a punch and die. The cutoff punch and die cut across the entire width of the strip. Frequently, an irregularly shaped cutoff operation may simultaneously give the workpiece all or part of the desired shape.

Dinking

Dinking is a modified shearing operation that is used to blank shapes from low-strength materials, primarily rubber, fiber, and cloth.

3.5 Drawing

Cold Drawing

Cold drawing is a term that can refer to two somewhat different operations. If the stock is in the form of sheet metal, cold drawing is the forming of parts wherein plastic flow occurs over a curved axis. This is one of the most important of all cold-working operations because a wide

range of parts, from small caps to large automobile body tops and fenders, can be drawn in a few seconds each. Cold drawing is similar to hot drawing, but the higher deformation forces, thinner metal, limited ductility, and closer dimensional tolerance create some distinctive problems.

If the stock is wire, rod, or tubing, cold drawing refers to the process of reducing the cross section of the material by pulling it through a die, a sort of tensile equivalent to extrusion.

Cold Spinning

Cold spinning is similar to hot spinning, discussed above.

Stretch Forming

In stretch forming, only a single male form block is required. The sheet of metal is gripped by two or more sets of jaws that stretch it and wrap it around the form block as the latter raises upward. Various combinations of stretching, wrapping, and upward motion of the blocks are used, depending on the shape of the part.

Shell or Deep Drawing

The drawing of closed cylindrical or rectangular containers, or a variation of these shapes, with a depth frequently greater than the narrower dimension of their opening, is one of the most important and widely used manufacturing processes. Because the process had its earliest uses in manufacturing artillery shells and cartridge cases, it is sometimes called *shell drawing*. When the depth of the drawn part is less than the diameter, or minimum surface dimension, of the blank, the process is considered to be *shallow drawing*. If the depth is greater than the diameter, it is considered to be *deep drawing*.

The design of complex parts that are to be drawn has been aided considerably by computer techniques but is far from being completely and successfully solved. Consequently, such design still involves a mix of science, experience, empirical data, and actual experimentation. The body of known information is quite substantial, however, and is being used with outstanding results.

Forming with Rubber or Fluid Pressure

Several methods of forming use rubber or fluid pressure (Fig. 16) to obtain the desired information and thereby eliminate either the male or female member of the die set. Blanks of sheet metal are placed on top of form blocks, which usually are made of wood. The upper ram, which contains a pad of rubber 8–10 in. (200–250 mm) thick in a steel container, then descends. The rubber pad is confined and transmits force to the metal, causing it to bend to the desired shape. Since no female die is used and form blocks replace the male die, die cost is quite low.

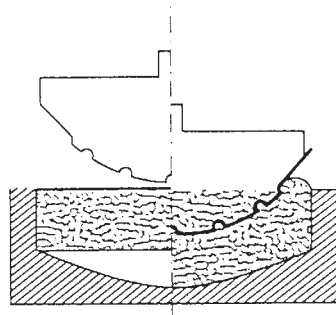


Figure 16 Form with rubber.

The hydroform process or “rubber bag forming” replaces the rubber pad with a flexible diaphragm backed by controlled hydraulic pressure. Deeper parts can be formed with truly uniform fluid pressure.

The bulging oil or rubber is used for applying an internal bulging force to expand a metal blank or tube outward against a female mold or die, thereby eliminating the necessity for a complicated, multiple-piece male die member.

Ironing

Ironing is the name given to the process of thinning the walls of a drawn cylinder by passing it between a punch and a die where the separation is less than the original wall thickness. The walls are elongated and thinned while the base remains unchanged. The most common example of an ironed product is the thin-walled all-aluminum beverage can.

Embossing

Embossing is a method for producing lettering or other designs in thin sheet metal. Basically, it is a very shallow drawing operation, usually in open dies, with the depth of the draw being from one to three times the thickness of the metal.

High-Energy-Rate Forming

A number of methods have been developed for forming metals through the release and application of large amounts of energy in a very short interval (Fig. 17). These processes are called *high-energy-rate-forming (HERF) processes*. Many metals tend to deform more readily under the ultrarapid rates of load application used in these processes, a phenomenon apparently related to the relative rates of load application and the movement of dislocations through the metal. As a consequence, HERF makes it possible to form large workpieces and difficult-to-form metals with less expensive equipment and tooling than would otherwise be required.

The high-energy release rates are obtained by five methods:

1. Underwater explosions
2. Underwater spark discharge (electrohydraulic techniques)
3. Pneumatic–mechanical means
4. Internal combustion of gaseous mixtures
5. Rapidly formed magnetic fields (electromagnetic techniques)

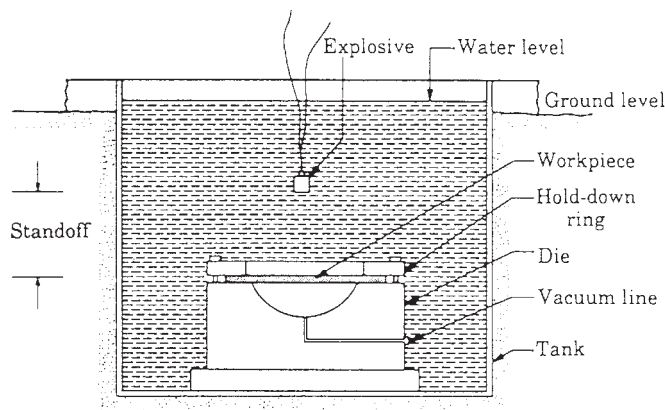


Figure 17 High-energy-rate forming.

4 METAL CASTING AND MOLDING PROCESSES

Casting provides versatility and flexibility that have maintained casting position as a primary production method for machine elements. Casting processes are divided according to the specific type of molding method used in casting, as follows:

1. Sand
2. Centrifugal
3. Permanent
4. Die
5. Plaster mold
6. Investment

4.1 Sand Casting

Sand casting consists basically of pouring molten metal into appropriate cavities formed in a sand mold (Fig. 18). The sand may be natural, synthetic, or an artificially blended material.

Molds

The two common types of sand molds are the *dry sand mold* and the *green sand mold*. In the dry sand mold, the mold is dried thoroughly prior to closing and pouring, while the green sand mold is used without any preliminary drying. Because the dry sand mold is more firm and resistant to collapse than the green sand mold, core pieces for molds are usually made in this way. Cores are placed in mold cavities to form the interior surfaces of castings.

Patterns

To produce a mold for a conventional sand cast part, it is necessary to make a pattern of the part. Patterns are made from wood or metal to suit a particular design, with allowances to compensate for such factors as natural metal shrinkage and contraction characteristics. These and other effects, such as mold resistance, distortion, casting design, and mold design, which are not entirely within the range of accurate prediction, generally make it necessary to adjust the pattern to produce castings of the required dimensions.

Access to the mold cavity for entry of the molten metal is provided by sprues, runners, and gates.

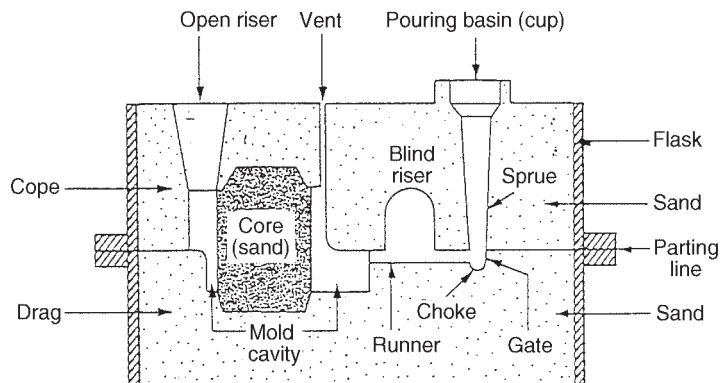


Figure 18 Sectional view of casting mold.

Shrinkage

Allowances must be made on patterns to counteract the contraction in size as the metal cools. The amount of shrinkage is dependent on the design of the coating, type of metal used, solidification temperature, and mold resistance. Table 3 gives average shrinkage allowance values used in sand casting. Smaller values apply generally to large or cored castings of intricate design. Larger values apply to small to medium simple castings designed with unrestrained shrinkage.

Machining

Allowances are required in many cases because of unavoidable surface impurities, warpage, and surface variations. Average machining allowances are given in Table 4. Good practice dictates

Table 3 Pattern Shrinkage Allowance (in./ft)

Metal	Shrinkage
Aluminum alloys	1/10–5/32
Beryllium copper	1/8–5/32
Copper alloys	3/16–7/32
Everdur	3/16
Gray irons	1/8
Hastelloy alloys	1/4
Magnesium alloys	1/8–11/64
Malleable irons	1/16–3/16
Meehanite	1/10–5/32
Nickel and nickel alloys	1/4
Steel	1/8–1/4
White irons	3/16–1/4

Table 4 Machining Allowances for Sand Castings (in.)

Metal	Casting Size	Finish Allowance
Cast irons	Up to 12 in.	3/32
	13–24 in.	1/8
	25–42 in.	3/16
	43–60 in.	1/4
	61–80 in.	5/16
	81–120 in.	3/8
Cast steels	Up to 12 in.	1/8
	13–24 in.	3/16
	25–42 in.	5/16
	43–60 in.	3/8
	61–80 in.	7/16
	81–120 in.	1/2
Malleable irons	Up to 8 in.	1/16
	9–12 in.	3/32
	13–24 in.	1/8
	25–36 in.	3/16
Nonferrous metals	Up to 12 in.	1/16
	13–24 in.	1/8
	25–36 in.	5/32

Table 5 Minimum Sections for Sand Castings (in.)

Metal	Section
Aluminum alloys	3/16
Copper alloys	3/32
Gray irons	1/8
Magnesium alloys	5/32
Malleable irons	1/8
Steels	1/4
White irons	1/8

use of minimum section thickness compatible with the design. The normal minimum section recommended for various metals is shown in Table 5.

4.2 Centrifugal Casting

Centrifugal casting consists of having a sand, metal, or ceramic mold that is rotated at high speeds. When the molten metal is poured into the mold, it is thrown against the mold wall, where it remains until it cools and solidifies. The process is increasingly being used for such products as cast-iron pipes, cylinder liners, gun barrels, pressure vessels, brake drums, gears, and flywheels. The metals used include almost all castable alloys. Most dental tooth caps are made by a combined lost-wax process and centrifugal casting.

Advantages and Limitations

Because of the relatively fast cooling time, centrifugal castings have a fine grain size. There is a tendency for the lighter nonmetallic inclusion, slag particles, and dross to segregate toward the inner radius of the castings (Fig. 19), where it can be easily removed by machining. Owing to the high purity of the outer skin, centrifugally cast pipes have a high resistance to atmospheric corrosion. Figure 19 shows a schematic sketch of how a pipe would be centrifugally cast in a horizontal mold. Parts that have diameters exceeding their length are produced by vertical-axis casting (see Fig. 20).

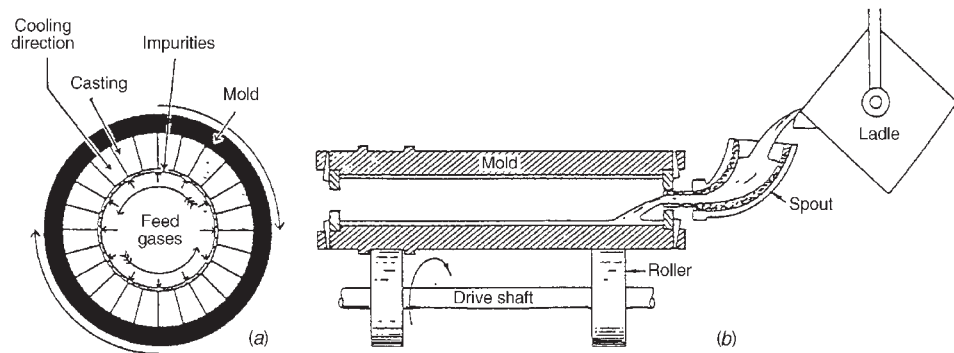


Figure 19 Principle of centrifugal casting is to produce the high-grade metal by throwing the heavier metal outward and forcing the impurities to congregate inward (a). Shown at (b) is a schematic of how a horizontal-bond centrifugal casting is made.

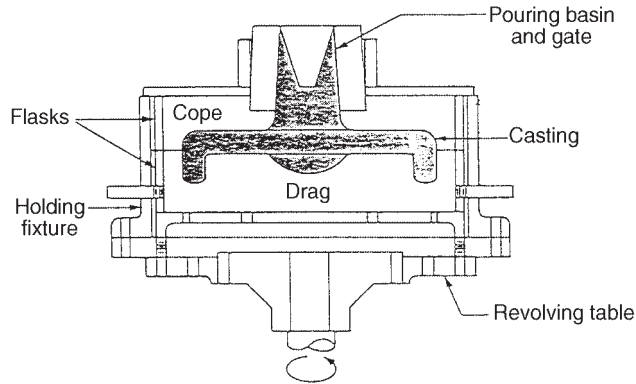


Figure 20 Floor-type vertical centrifugal casting machine for large-diameter parts.

If the centrifugal force is too low or too great, abnormalities will develop. Most horizontal castings are spun so that the force developed is about 65*g*. Vertically cast parts force is about 90–100*g*.

The centrifugal force (CF) is calculated from

$$CF = \frac{mv^2}{r} \text{ lb}$$

$$m = \text{mass} = \frac{W}{g} = \frac{\text{weight, lb}}{\text{acceleration of gravity (ft/s)}^2} = \frac{W}{32.2}$$

where $v = \text{velocity, ft/s} = r \times w$
 $r = \text{radius, ft} = 1/2D$
 $w = \text{angular velocity, rad/s}$
 $w = 2\pi/60 \times \text{rpm}$
 $D = \text{inside diameter, ft}$

The number of *g*'s is

$$g = \frac{CF}{W}$$

Hence,

$$g's = \frac{1}{W} \times \left[\frac{W}{32.2 \times r} \left(\frac{r \times 2\pi}{60} \right)^2 \right]$$

$$= r \times 3.41 \times 10^{-4} \text{ rpm}^2$$

$$= 1.7 \times 10^{-4} \times D \times (\text{rpm})^2$$

The spinning speed for horizontal-axis molds may be found in English units from the equation

$$N = \sqrt{(\text{number of } g's) \times \frac{70,500}{D}}$$

where $N = \text{rpm}$
 $D = \text{inside diameter of mold, ft}$

4.3 Permanent-Mold Casting

As demand for quality castings in production quantities increased, the attractive possibilities of metal molds brought about the development of the permanent-mold process. Although not as flexible regarding design as sand casting, metal-mold casting made possible the continuous production of quantities of casting from a single mold as compared to batch production of individual sand molds.

Metal Molds and Cores

In permanent-mold casting, both metal molds and cores are used, the metal being poured into the mold cavity with the usual gravity head as in sand casting. Molds are normally made of dense iron or meehanite, large cores of cast iron, and small or collapsible cores of alloy steel. All necessary sprues, runners, gates, and risers must be machined into the mold, and the mold cavity itself is made with the usual metal shrinkage allowances. The mold is usually composed of one, two, or more parts, which may swing or slide for rapid operation. Whereas in sand casting the longest dimension is always placed in a horizontal position, in permanent-mold casting the longest dimension of a part is normally placed in a vertical position.

Production Quantities

Wherever quantities are in the range of 500 pieces or more, permanent-mold casting becomes competitive in cost with sand casting, and if the design is simple, runs as small as 200 pieces are often economical. Production runs of 1000 pieces or more will generally produce a favorable cost difference. High rates of production are possible, and multiple-cavity dies with as many as 16 cavities can be used. In casting gray iron in multiple molds, as many as 50,000 castings per cavity are common with small parts. With larger parts of gray iron, weighing from 12–15 lb, single-cavity molds normally yield 2000–3000 pieces per mold on an average. Up to 100,000 parts per cavity or more are not uncommon with nonferrous metals, magnesium providing the longest die life. Low-pressure permanent mold casting is economical for quantities up to 40,000 pieces (Fig. 21).

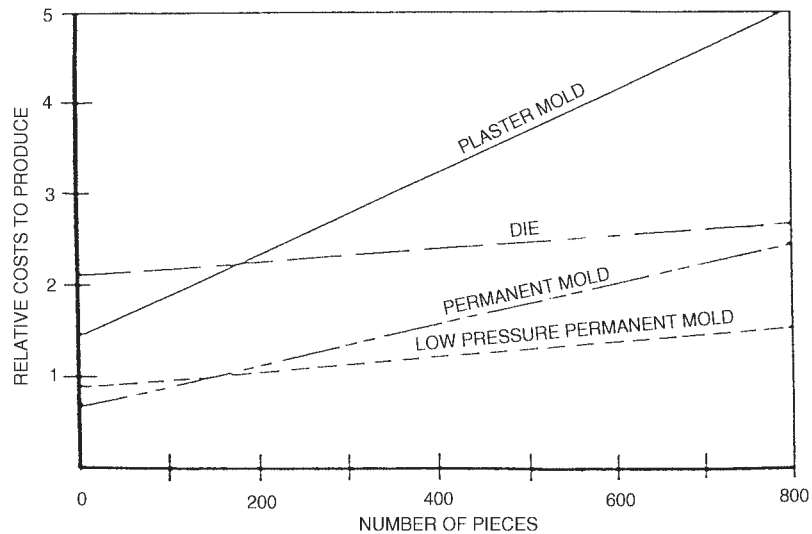


Figure 21 Cost comparison of various casting systems.

Die Casting

Die casting may be classified as a permanent-mold casting system; however, it differs from the process just described in that the molten metal is forced into the mold or die under high pressure [1000–30,000 psi (6.89–206.8 MPa)]. The metal solidifies rapidly (within a fraction of a second) because the die is water cooled. Upon solidification, the die is opened and ejector pins automatically knock the casting out of the die. If the parts are small, several of them may be made at one time in what is termed a *multicavity die*.

There are two main types of machines used: the hot-chamber and the cold-chamber types.

Hot-Chamber Die Casting. In the hot-chamber machine, the metal is kept in a heated holding pot. As the plunger descends, the required amount of alloy is automatically forced into the die. As the piston retracts, the cylinder is again filled with the right amount of molten metal. Metals such as aluminum, magnesium, and copper tend to alloy with the steel plunger and cannot be used in the hot chamber.

Cold-Chamber Die Casting. This process gets its name from the fact that the metal is ladled into the cold chamber for each shot. This procedure is necessary to keep the molten-metal contact time with the steel cylinder to a minimum. Iron pickup is prevented, as is freezing of the plunger in the cylinder.

Advantages and Limitations

Die-casting machines can produce large quantities of parts with close tolerances and smooth surfaces. The size is limited only by the capacity of the machine. Most die castings are limited to about 75 lb (34 kg) of zinc, 65 lb (30 kg) of aluminum, and 44 lb (20 kg) of magnesium. Die castings can provide thinner sections than any other casting process. Wall thickness as thin as 0.015 in. (0.38 mm) can be achieved with aluminum in small items. However, a more common range on larger sizes will be 0.105–0.180 in. (2.67–4.57 mm).

Some difficulty is experienced in getting sound castings in the larger capacities. Gases tend to be entrapped, which results in low strength and annoying leaks. Of course, one way to reduce metal sections without sacrificing strength is to design in ribs and bosses. Another approach to the porosity problem has been to operate the machine under vacuum. This process is now being developed.

The surface quality is dependent on that of the mold. Parts made from new or repolished dies may have a surface roughness of 24 $\mu\text{in.}$ (0.61 μm). The high surface finish available means that, in most cases, coatings such as chromeplating, anodizing, and painting may be applied directly. More recently, decorative finishes of texture, as obtained by photoetching, have been applied. The technique has been used to simulate wood-grain finishes, as well as textile and leather finishes, and to obtain checkering and cross-hatching.

4.4 Plaster Mold Casting

In general, the various methods of plaster mold casting are similar. The plaster, also known as *gypsum* or *calcium sulfate*, is mixed dry with other elements, such as talc, sand, asbestos, and sodium silicate. To this mix is added a controlled amount of water to provide the desired permeability in the mold. The slurry that results is heated and delivered through a hose to the flasks, all surfaces of which have been sprayed with a parting compound. The plaster slurry readily fills in and around the most minute details in the highly polished brass patterns. Following filling, the molds are subjected to a short period of vibration and the slurry sets in 5–10 min.

Molds

Molds are extracted from the flask with a vacuum head, following which drying is completed in a continuous oven. Copes and drags are then assembled, with cores when required, and the

224 Metal Forming, Shaping, and Casting

castings are poured. Upon solidification, the plaster is broken away and any cores used are washed out with a high-pressure jet of water.

4.5 Investment Casting

Casting processes in which the pattern is used only once are variously referred to as *lost-wax* or *precision-casting* processes. They involve making a pattern of the desired form out of wax or plastic (usually polystyrene). The expendable pattern may be made by pressing the wax into a split mold or by the use of an injection-molding machine. The patterns may be gated together so that several parts can be made at once. A metal flask is placed around the assembled patterns and a refractory mold slurry is poured in to support the patterns and form the cavities. A vibrating table equipped with a vacuum pump is used to eliminate all the air from the mold. Formerly, the standard procedure was to dip the patterns in the slurry several times until a coat was built up. This is called the *investment process*. After the mold material has set and dried, the pattern material is melted and allowed to run out of the mold.

The completed flasks are heated slowly to dry the mold and to melt out the wax, plastic, or whatever pattern material was used. When the molds have reached a temperature of 100°F (37.8°C), they are ready for pouring. Vacuum may be applied to the flasks to ensure complete filling of the mold cavities.

When the metal has cooled, the investment material is removed by vibrating hammers or by tumbling. As with other castings, the gates and risers are cut off and ground down.

Ceramic Process

The ceramic process is somewhat similar to the investment casting in that a creamy ceramic slurry is poured over a pattern. In this case, however, the pattern, made out of plastic, plaster, wood, metal, or rubber, is reusable. The slurry hardens on the pattern almost immediately and becomes a strong green ceramic of the consistency of vulcanized rubber. It is lifted off the pattern while it is still in the rubberlike phase. The mold is ignited with a torch to burn off the volatile portion of the mix. It is then put in a furnace and baked at 1800°F (982°C), resulting in a rigid refractory mold. The mold can be poured while still hot.

Full-Mold Casting

Full-mold casting may be considered a cross between conventional sand casting and the investment technique of using lost wax. In this case, instead of a conventional pattern of wood, metals, or plaster, a polystyrene foam or styrofoam is used. The pattern is left in the mold and is vaporized by the molten metal as it rises in the mold during pouring. Before molding, the pattern is usually coated with a zirconite wash in an alcohol vehicle. The wash produces a relatively tough skin separating the metal from the sand during pouring and cooling. Conventional foundry sand is used in backing up the mold.

5 PLASTIC MOLDING PROCESSES

Plastic molding is similar in many ways to metal molding. For most molding operations, plastics are heated to a liquid or a semifluid state and are formed in a mold under pressure. Some of the most common molding processes are discussed below.

5.1 Injection Molding

The largest quantity of plastic parts is made by injection molding. Plastic compound is fed in powdered or granular form from a hopper through metering and melting stages and then injected into a mold. After a brief cooling period, the mold is opened and the solidified part is ejected.

5.2 Coinjection Molding

Coinjection molding makes it possible to mold articles with a solid skin of one thermoplastic and a core of another thermoplastic. The skin material is usually solid while the core material contains blowing agents.

The basic process may be one-, two-, or three-channel technology. In one-channel technology, the two melts are injected into the mold, one after the other. The skin material cools and adheres to the colder surface; a dense skin is formed under proper parameter settings. The thickness of the skin can be controlled by adjustment of injection speed, stock temperature, mold temperature, and flow compatibility of the two melts.

In two- and three-channel techniques, both plastic melts may be introduced simultaneously. This allows for better control of wall thickness of the skin, especially in gate areas on both sides of the part.

Injection-Molded Carbon Fiber Composites

By mixing carbon or glass fibers in injection-molded plastic parts, they can be made lightweight yet stiffer than steel.

5.3 Rotomolding

In rotational molding, the product is formed inside a closed mold that is rotated about two axes as heat is applied. Liquid or powdered thermoplastic or thermosetting plastic is poured into the mold, either manually or automatically.

5.4 Expandable-Bead Molding

The expandable-bead process consists of placing small beads of polystyrene along with a small amount of blowing agent in a tumbling container. The polystyrene beads soften under heat, which allows a blowing agent to expand them. When the beads reach a given size, depending on the density required, they are quickly cooled. This solidifies the polystyrene in its larger foamed size. The expanded beads are then placed in a mold until it is completely filled. The entrance port is then closed and steam is injected, resoftening the beads and fusing them together. After cooling, the finished, expanded part is removed from the mold.

5.5 Extruding

Plastic extrusion is similar to metal extrusion in that a hot material (plastic melt) is forced through a die having an opening shaped to produce a desired cross section. Depending on the material used, the barrel is heated anywhere from 250 to 600°F (121 – 316°C) to transform the thermoplastic from a solid to a melt. At the end of the extruder barrel is a screen pack for filtering and building back pressure. A breaker plate serves to hold the screen pack in place and straighten the helical flow as it comes off the screen.

5.6 Blow Molding

Blow molding is used extensively to make bottles and other lightweight, hollow plastic parts. Two methods are used: injection blow molding and extrusion blow molding.

Injection blow molding is used primarily for small containers. The parison (molten-plastic pipe) or tube is formed by the injection of plasticized material around a hollow mandrel. While the material is still molten and still on the mandrel, it is transferred into the blowing mold where air is used to inflate it. Accurate threads may be formed at the neck.

In extrusion-type blow molding, parison is inflated under relatively low pressure inside a split-metal mold. The die closes, pinching the end and closing the top around the mandrel. Air enters through the mandrel and inflates the tube until the plastic contacts the cold wall, where it solidifies. The mold opens, the bottle is ejected, and the tailpiece falls off.

5.7 Thermoforming

Thermoforming refers to heating a sheet of plastic material until it becomes soft and pliable and then forming it either under vacuum, by air pressure, or between matching mold halves.

5.8 Reinforced-Plastic Molding

Reinforced plastics generally refers to polymers that have been reinforced with glass fibers. Other materials used are asbestos, sisal, synthetic fibers such as nylon and polyvinyl chloride, and cotton fibers. High-strength composites using graphite fibers are now commercially available with moduli of 50,000,000 psi (344,700,000 MPa) and tensile strengths of about 300,000 psi (2,068,000 MPa). They are as strong as or stronger than the best alloy steels and are lighter than aluminum.

5.9 Forged-Plastic Parts

The forging of plastic materials is a relatively new process. It was developed to shape materials that are difficult or impossible to mold and is used as a low-cost solution for small production runs.

The forging operation starts with a blank or billet of the required shape and volume for the finished part. The blank is heated to a preselected temperature and transferred to the forging dies, which are closed to deform the work material and fill the die cavity. The dies are kept in the closed position for a definite period of time, usually 15–60 s. When the dies are opened, the finished forging is removed. Since forging involves deformation of the work material in a heated and softened condition, the process is applicable only to thermoplastics.

6 POWDER METALLURGY

In powder metallurgy (PM), fine metal powders are pressed into a desired shape, usually in a metal die and under high pressure, and the compacted powder is then heated (sintered), with a protective atmosphere. The density of sintered compacts may be increased by repressing. Repressing is also performed to improve the dimensional accuracy, either concurrently or subsequently, for a period of time at a temperature below the melting point of the major constituent. PM has a number of distinct advantages that account for its rapid growth in recent years, including (1) no material is wasted, (2) usually no machining is required, (3) only semiskilled labor is required, and (4) some unique properties can be obtained, such as controlled degrees of porosity and built-in lubrication.

A crude form of PM appears to have existed in Egypt as early as 3000 BC, using particles of sponge iron. In the nineteenth century, PM was used for producing platinum and tungsten wires. However, its first significant use related to general manufacturing was in Germany, following World War I, for making tungsten carbide cutting-tool tips. Since 1945 the process has been highly developed, and large quantities of a wide variety of PM products are made annually, many of which could not be made by any other process. Most are under 2 in. (50.8 mm) in size, but many are larger, some weighing up to 50 lb (22.7 kg) and measuring up to 20 in. (508 mm).

Powder metallurgy normally consists of four basic steps:

1. Producing a fine metallic powder
2. Mixing and preparing the powder for use
3. Pressing the powder into the desired shape
4. Heating (sintering) the shape at an elevated temperature

Other operations can be added to obtain special results.

The pressing and sintering operations are of special importance. The pressing and repressing greatly affect the density of the product, which has a direct relationship to the strength properties. Sintering strips contaminants from the surface of the powder particles, permitting diffusion bonding to occur and resulting in a single piece of material. Sintering usually is done in a controlled, inert atmosphere, but sometimes it is done by the discharge of spark through the powder while it is under compaction in the mold.

6.1 Properties of PM Products

Because the strength properties of PM products depend on so many variables—type and size of powder, pressing pressure, sintering temperature, finishing treatments, and so on—it is difficult to give generalized information. In general, the strength properties of products that are made from pure metals (unalloyed) are about the same as those made from the same wrought metals. As alloying elements are added, the resulting strength properties of PM products fall below those of wrought products by varying, but usually substantial, amounts. The ductility usually is markedly less, as might be expected because of the lower density. However, tensile strengths of 40,000–50,000 psi (275.8–344.8 MPa) are common, and strengths above 100,000 psi (689.5 MPa) can be obtained. As larger presses and forging combined with PM preforms are used, to provide greater density, the strength properties of PM materials will more nearly equal those of wrought materials. Coining can also be used to increase the strength properties of PM products and to improve their dimensional accuracy.

7 SURFACE TREATMENT

Products that have been completed to their proper shape and size frequently require some type of surface finishing to enable them to satisfactorily fulfill their function. In some cases, it is necessary to improve the physical properties of the surface material for resistance to penetration or abrasion.

Surface finishing may sometimes become an intermediate step in processing. For instance, cleaning and polishing are usually essential before any kind of plating process. Another important need for surface finishing is for corrosion protection in a variety of environments. The type of protection provided will depend largely on the anticipated exposure, with due consideration to the material being protected and the economic factors involved.

Satisfying the above objectives necessitates the use of many surface-finishing methods that involve chemical change of the surface; mechanical work affecting surface properties, cleaning by a variety of methods; and the application of protective coatings organic and metallic.

7.1 Cleaning

Few, if any, shaping and sizing processes produce products that are usable without some type of cleaning unless special precautions are taken. Figure 22 indicates some of the cleaning methods available. Some cleaning methods provide multiple benefits. Cleaning and finish improvements are often combined. Probably of even greater importance is the combination of corrosion protection with finish improvement, although corrosion protection is more often a second step that involves coating an already cleaned surface with some other material or chemical conversion.

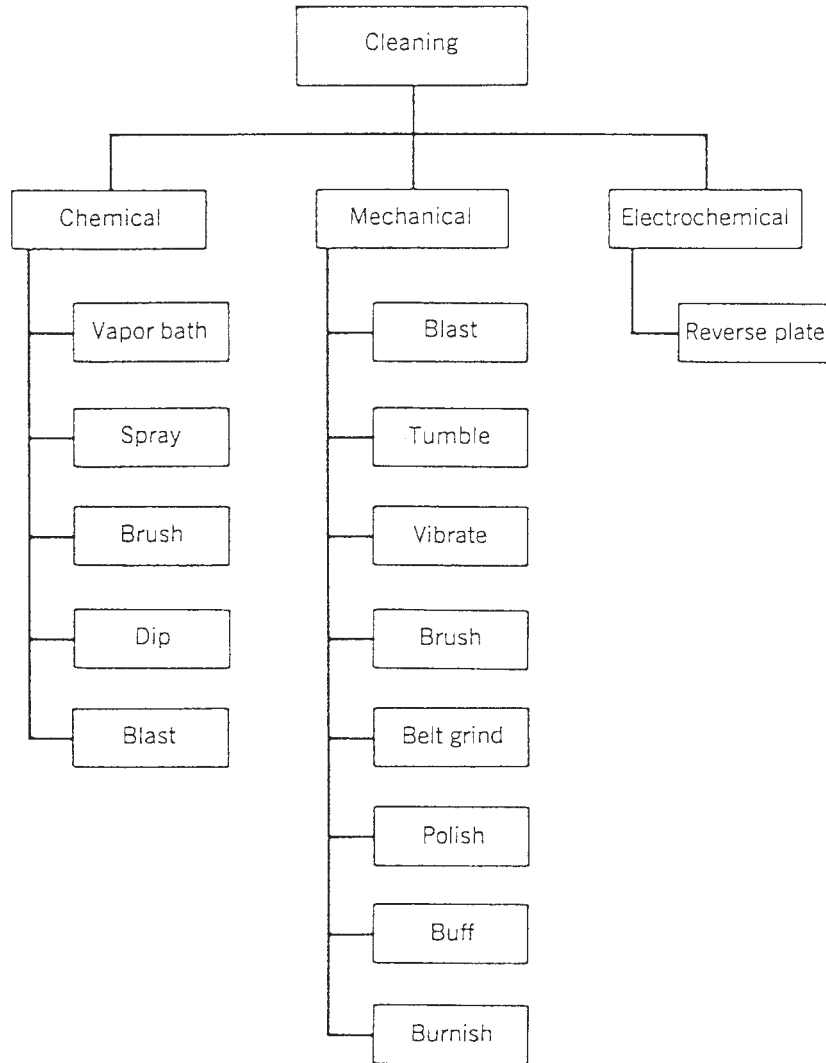


Figure 22 Cleaning methods.

Liquid and Vapor Baths

Liquid and Vapor Solvents. The most widely used cleaning methods make use of a cleaning medium in liquid or vapor form. These methods depend on a solvent or chemical action between the surface contaminants and the cleaning material.

Petroleum Solvents. Among the more common cleaning jobs required is the removal of grease and oil deposited during manufacturing or intentionally coated on the work to provide protection. One of the most efficient ways to remove this material is by use of solvents that dissolve the grease and oil but have no effect on the base metal. Petroleum derivatives, such as Stoddard solvent and kerosene, are common for this purpose, but, since they introduce some

danger of fire, chlorinated solvents, such as trichlorethylene, that are free of this fault are sometimes substituted.

Conditioned Water. One of the most economical cleaning materials is water. However, it is seldom used alone, even if the contaminant is fully water soluble, because the impurity of the water itself may contaminate the work surface. Depending on its use, water is treated with various acids and alkalis to suit the job being performed.

Pickling. Water containing sulfuric acid in a concentration from about 10 to 25% and at a temperature of approximately 149°F (65°C) is commonly used in a process called *pickling* for removal of surface oxides or scale or iron and steel.

Mechanical Work Frequently Combined with Chemical Action. Spraying, brushing, and dipping methods are also used with liquid cleaners. In nearly all cases, mechanical work to cause surface film breakdown and particle movement is combined with chemical and solvent action. The mechanical work may be agitation of the product, as in dipping, movement of the cleaning agent, as in spraying, or use of a third element, as in rubbing solidus brushing. In some applications, sonic or ultrasonic vibrations are applied to either the solution or the workpieces to speed the cleaning action. Chemical activity is increased with higher temperatures and optimum concentration of the cleaning agent, both of which must in some cases be controlled closely for efficient action.

Blasting

The term *blasting* is used to refer to all those cleaning methods in which the cleaning medium is accelerated to high velocity and impinged against the surface to be cleaned. The high velocity may be provided by air or water directed through a nozzle or by mechanical means with a revolving slinger. The cleaning agent may be either dry or wet solid media, such as sand, abrasive, steel grit, or shot, or may be liquid or vapor solvents combined with abrasive material. In addition to cleaning, solid particles can improve finish and surface properties of the material on which they are used. Blasting tends to increase the surface area and thus set up compressive stresses that may cause a warping of thin sections, but in other cases, it may be very beneficial in reducing the likelihood of fatigue failure. When used for the latter purpose, the process is more commonly known as *shot peening*.

Water Slurries. Liquid or vaporized solvents may, by themselves, be blasted against a surface for high-speed cleaning of oil and grease films with both chemical and mechanical action. Water containing rust-inhibiting chemicals may carry, in suspension, fine abrasive particles that provide a grinding cutting-type action for finish improvement along with cleaning. The blasting method using this medium is commonly known as *liquid honing*.

Abrasive Barrel Finishing

Barrel finishing, rolling, tumbling, and rattling are terms used to describe similar operations that consist of packing parts together with some cleaning media in a cylinder or drum, which can be rotated to cause movement among them. The media may be abrasive (either fine or coarse): metal stars, slugs, or balls; stones; wood chips; sawdust; or cereals. The work may be done wet or dry, depending on the materials being worked with, the kind of surface finish desired, and the kind of equipment available.

Wire Brushing

A number of cleaning operations can be quickly and easily performed by use of a high-speed rotating wire brush. In addition to cleaning, the contact rubbing of the wire ends across the work surface produce surface improvement by a burnishing-type action. Sharp edges and burrs can be removed.

Abrasive Belt Finishing

Continuous fabric belts coated with abrasive can be driven in several kinds of machines to provide a straight-line cutting motion for grinding, smoothing, and polishing work surfaces. Plane surfaces are the most common surfaces worked on with fabric belts.

Polishing

The term *polishing* may be interpreted to mean any nonprecision procedure providing a glossy surface but is most commonly used to refer to a surface-finishing process using a flexible abrasive wheel. The wheels may be constructed of felt or rubber with an abrasive band, of multiple coated abrasive disks, of leaves of coated abrasive, of felt or fabric to which loose abrasive is added as needed, or of abrasives in a rubber matrix.

Buffing

About the only difference between buffing and polishing is that, for buffing, a fine abrasive carried in wax or a similar substance is charged on the surface of a flexible level.

Electropolishing

If a workpiece is suspended in an electrolyte and connected to the anode in an electrical circuit, it will supply metal to the electrolyte in a reverse plating process. Material will be removed faster from the high spots of the surface than from the depressions and will thereby increase the average smoothness. The cost of the process is prohibitive for very rough surfaces because larger amounts of metal must be removed to improve surface finish than would be necessary for the same degree of improvement by mechanical polishing. Electropolishing is economical only for improving a surface that is already good or for polishing complex and irregular shapes, the surfaces of which are not accessible to mechanical polishing and buffing equipment.

7.2 Coatings

Many products, particularly those exposed to view and those subject to change by the environment with which they are in contact, need some type of coating for improved appearance or for protection from chemical attack. The need for corrosion protection for maintenance and appearance is important. In addition to change of appearance, loss of actual material, change of dimensions, and decrease of strength, corrosion may be the cause of eventual loss of service or failure of a product. Material that must carry loads in structural applications, especially when the loads are cyclic in nature, may fail with fatigue if corrosion is allowed to take place. Corrosion occurs more readily in highly stressed material, where it attacks grain boundaries in such a way as to form points of stress concentration that may be nuclei for fatigue failure.

Harness and wear resistance, however, can be provided on a surface by plating with hard metals. Chromium plating of gauges and other parts subject to abrasion is frequently used to increase their wear life. Coatings of plastic material and asphaltic mixtures are sometimes placed on surfaces to provide sound deadening. The additional benefit of protection from corrosion is usually acquired at the same time.

Plastics of many kinds, mostly of the thermoplastic type because they are easier to apply and also easier to remove later if necessary, are used for mechanical protection. Highly polished material may be coated with plastic, which may be stripped off later, to prevent abrasion and scratches during processing. It is common practice to coat newly sharpened cutting edges of tools by dipping them in thermoplastic material to provide mechanical protection during handling and storage.

Organic Coatings

Organic coatings are used to provide pleasing colors, to smooth surfaces, to provide uniformity in both color and texture, and to act as a protective film for control of corrosion. Organic resin coatings do not ordinarily supply any chemical-inhibiting qualities. Instead, they merely provide a separating film between the surface to be protected and the corrosive environment. The important properties, therefore, are continuity, permeability, and adhesion characteristics.

Paints, Varnishes, and Enamels

Paints. Painting is a generic term that has come to mean the application of almost any kind of organic coating by any method. Because of this interpretation, it is also used generally to describe a broad class of products. As originally defined and as used most at present, paint is a mixture of pigment in a drying oil. The oil serves as a carrier for the pigment and in addition creates a tough continuous film as it dries. Drying oils, one of the common ones of which is linseed oil, become solid when large surface areas are exposed to air. Drying starts with a chemical reaction of oxidation. Nonreversible polymerization accompanies oxidation to complete the change from liquid to solid.

Varnish. Varnish is a combination of natural or synthetic resins and drying oil, sometimes containing volatile solvents as well. The material dries by a chemical reaction in the drying oil to a clear or slightly amber-colored film.

Enamel. Enamel is a mixture of pigment in varnish. The resins in the varnish cause the material to dry to a smoother, harder, and glossier surface than is produced by ordinary paints. Some enamels are made with thermosetting resins that must be baked for complete dryness. These baking enamels provide a toughness and durability not usually available with ordinary paints and enamels.

Lacquers

The term *lacquer* is used to refer to finishes consisting of thermoplastic materials dissolved in fast-drying solvents. One common combination is cellulose nitrate dissolved in butyl acetate. Present-day lacquers are strictly air-drying and form films very quickly after being applied, usually by spraying. No chemical change occurs during the hardening of lacquers; consequently, the dry film can be redissolved in the thinner. Cellulose acetate is used in place of cellulose nitrate in some lacquers because it is nonflammable. Vinyls, chlorinated hydrocarbons, acrylics, and other synthetic thermoplastic resins are also used in the manufacture of lacquers.

Vitreous Enamels

Vitreous, or porcelain, enamel is actually a thin layer of glass fused onto the surface of a metal, usually steel or iron. Shattered glass, ball milled in a fine particle size, is called *frit*. Frit is mixed with clay, water, and metal oxides, which produce the desired color, to form a thin slurry called *slip*. This is applied to the prepared metal surface by dipping or spraying and, after drying, is fired at approximately 1470°F (800°C) to fuse the material to the metal surface.

Metallizing

Metal spraying, or metallizing, is a process in which metal wire or powder is fed into an oxy-acetylene heating flame and then, after melting, is carried by high-velocity air to be impinged against the work surface. The small droplets adhere to the surface and bond together to build up a coating.

Vacuum Metallizing

Some metals can be deposited in very thin films, usually for reflective or decorative purposes, as a vapor deposit. The metal is vaporized in a high-vacuum chamber containing the parts to

be coated. The metal vapor condenses on the exposed surfaces in a thin film that follows the surface pattern. The process is cheap for coating small parts, considering the time element only, but the cost of special equipment needed is relatively high.

Aluminum is the most used metal for deposit by this method and is used frequently for decorating or producing a mirror surface on plastics. The thin films usually require mechanical protection by covering with lacquer or some other coating material.

Hot-Dip Plating

Several metals, mainly zinc, tin, and lead, are applied to steel for corrosion protection by a hot-dip process. Steel in sheet, rod, pipe, or fabricated form, properly cleansed and fluxed, is immersed in molten plating metal. As the work is withdrawn, the molten metal that adheres solidifies to form a protective coat. In some of the large mills, the application is made continuously to coil stock that is fed through the necessary baths and even finally inspected before being recoiled or cut into sheets.

Electroplating

Coatings of many metals can be deposited on other metals, and on nonmetals when suitably prepared, by electroplating. The objectives of plating are to provide protection against corrosion, to improve appearance, to establish wear- and abrasion-resistant surfaces, to add material for dimensional increase, and to serve as an intermediate step of multiple coating. Some of the most common metals deposited in this way are copper, nickel, cadmium, zinc, tin, silver, and gold. The majority are used to provide some kind of corrosion protection but appearance also plays a strong part in their use.

Temporary Corrosion Protection

It is not uncommon in industry for periods of time, sometimes quite long periods, to elapse between manufacture, assembly, shipment, and use of parts. Unless a new processing schedule can be worked out, about the only cure for the problem is corrosion protection suitable for the storage time and exposure. The coatings used are usually nondrying organic materials, called *shushing compounds*, that can be removed easily. The two principal types of compounds used for this purpose are petroleum-based materials, varying from extremely light oils to semisolids, and thermoplastics. The most common method of application of shushing compounds for small parts is by dipping. Larger parts that cannot be handled easily may be sprayed, brushed, or flow coated with the compound.

7.3 Chemical Conversions

A relatively simple and often fully satisfactory method for protection from corrosion is by conversion of some of the surface material to a chemical composition that resists from the environment. These converted metal surfaces consist of relatively thin [seldom more than 0.001 in. (0.025 mm) thick] inorganic films that are formed by chemical reaction with the base material. One important feature of the conversion process is that the coatings have little effect on the product dimensions.

Anodizing

Aluminum, magnesium, and zinc can be treated electrically in a suitable electrolyte to produce a corrosion-resistant oxide coating. The metal being treated is connected to the anode in the circuit, which provides the name *anodizing* for the process. Aluminum is commonly treated by anodizing that produces an oxide film thicker than, but similar to, that formed naturally with

exposure to air. Anodizing of zinc has very limited use. The coating produced on magnesium is not as protective as that formed on aluminum, but does provide some protective value and substantially increases protection when used in combination with paint coatings.

Chromate Coatings

Zinc is usually considered to have relatively good corrosion resistance. This is true when the exposure is to normal outdoor atmosphere where a relatively thin corrosion film forms. Contact with either highly aerated water films or immersion in stagnant water containing little oxygen causes uneven corrosion and pitting. The corrosion products of zinc are less dense than the base material, so that heavy corrosion not only destroys the product appearance, but also may cause malfunction by binding moving parts. Corrosion of zinc can be substantially slowed by the production of chromium salts on its surface. The corrosion resistance of magnesium alloys can be increased by immersion of anodic treatment in acid baths containing dichromates. Chromate treatment of both zinc and magnesium improves corrosion resistance but is used also to improve adhesion of paint.

Phosphate Coatings

Phosphate coatings, used mostly on steel, result from a chemical reaction of phosphoric acid with the metal to form a nonmetallic coating that is essentially phosphoric salts. The coating is produced by immersing small items or spraying large items with the phosphating solution. Phosphate surfaces may be used alone for corrosion resistance, but their most common application is as a base for paint coatings. Two of the most common application methods are called *parkerizing* and *bonderizing*.

Chemical Oxide Coatings

A number of proprietary blacking processes, used mainly on steel, produce attractive black oxide coatings. Most of the processes involve the immersing of steel in a caustic soda solution, heated to about 300°F (150°C) and made strongly oxidizing by the addition of nitrites or nitrates. Corrosion resistance is rather poor unless improved by application of oil, lacquer, or wax. As in the case of most of the other chemical conversion procedures, this procedure also finds use as a base for paint finishes.

BIBLIOGRAPHY

- Abrasion-Resistant Cast Iron Handbook*, American Foundry Society, 2000.
- T. Altan, S.-I. Oh, and H. Gegel, *Metal Forming—Fundamentals and Application*, ASM International, Materials Park, OH, 1983.
- ASM Handbook*, Vol. 14: *Forming and Forging*, ASM International, Materials Park, OH, 1988.
- B. Bhushan, and B. K. Gupta, *Handbook of Tribology: Materials, Coating, and Surface Treatment*, McGraw-Hill, New York, 1991.
- B. Bhushan, *Modern Tribology Handbook*, CRC Press, Boca Raton, FL, 2001.
- A. J. Clegg, *Precision Casting Processes*, Pergamon, New York, 1991.
- V. B. Ginzburg, *High-Quality Steel Rolling: Theory and Practice*, Dekker, New York, 1993.
- H. Hoffman, *Metal Forming Handbook*, Springer, 1998.
- W. F. Hosford, and R. M. Caddell, *Metal Forming, Mechanics and Metallurgy*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1993.
- Investment Casting Handbook*, Investment Casting Institute, 1997.
- K. Lange, *Handbook of Metal Forming*, McGraw-Hill, New York, 1985.
- J. H. Lindsay, *Coatings and Coating Processes for Metals*, ASM International, Materials Park, OH, 1998.

- Z. Marciniak, and J. L. Duncan, *The Mechanics of Sheet Metal Forming*, Edward Arnold, 1992.
- I. Suchy, *Handbook of Die Design*, McGraw-Hill, New York, 1997.
- Tool and Manufacturing Engineers Handbook*, 4th ed., Vol. 2: *Forming*, Society of Manufacturing Engineers, Dearborn, MI, 1984.
- B. Upton, *Pressure Die Casting*, Part 1: *Metals, Machines, Furnaces*, Pergamon, New York, 1982.
- R. H. Wagoner, and J. L. Chenot, *Fundamentals of Metal Forming*, Wiley, New York, 1996.
- R. H. Wagoner, and J. L. Chenot, *Metal Forming Analysis*, Cambridge University Press, Cambridge, UK, 2001.
- C. F. Walton, and T. J. Opar, *Iron Castings Handbook*, 3rd ed., Iron Castings Society, 1981.
- P. P. Wieser, *Steel Castings Handbook*, 6th ed., ASM International, Materials Park, OH, 1995.
- K. P. Young, *Semi-Solid Processing*, Chapman & Hall, 1997.
- K.-O. Yu, *Modeling for Casting and Solidification Processing*, Dekker, New York, 2001.
- M. E. Zohdi, "Statistical Analysis: Estimation and Optimization of Surface Finish," Proceedings of International Conference on Development of Production Systems, Copenhagen, Denmark, 1974.

CHAPTER 7

COATINGS AND SURFACE ENGINEERING: PHYSICAL VAPOR DEPOSITION

Allan Matthews
Sheffield University, Sheffield, United Kingdom

Suzanne L. Rohde
Infinidium, LLC, Steamboat Springs, Colorado

1	INTRODUCTION	235	4.1	Evaporative Processes	239
2	GLOW DISCHARGE PLASMA	236	4.2	Sputter Deposition Processes	242
3	FILM FORMATION AND GROWTH	237	4.3	Beam Processes	247
4	PROCESS DETAILS	239	5	FINAL COMMENTS	249
				REFERENCES	249

1 INTRODUCTION

The term *physical vapor deposition* (PVD) is used to describe processes in which at least one of the coating species is atomized from a solid source within a coating chamber, to then condense on a substrate, forming a film. It is different from *chemical vapor deposition* (CVD), as that process utilizes gaseous reagents as the source of coating material. Also, PVD is typically carried out under low-pressure vacuum conditions, whereas CVD can be performed over a wide range of operating pressures, from high vacuum to atmospheric.

PVD processes can be categorized according to the means of atomizing the source material; the main division, as shown in Fig. 1a, is between evaporative methods and sputtering. Evaporation, as the name suggests, involves the thermal vaporization of the source, whereas sputtering is a kinetically controlled mechanism in which the source material, or “target,” is bombarded with gas atoms, which then transfer momentum to atoms in the target, leading to the ejection of coating material atoms. The sputtering and evaporation techniques both originated at about the same time. The first sputtering experiments were reported by Grove¹ in 1852 and the first reports of evaporation deposition were by Faraday² in 1857.

Although PVD was originally used as a means of depositing elemental metallic coatings, it has been increasingly used for alloy and ceramic deposition. In the latter case this can be achieved by using a ceramic source or a metal source and inletting a reactive gas such as oxygen, nitrogen, or methane, for example, to produce oxides, nitrides, or carbides. Similarly, different means now exist to deposit multielement alloy or even composite metal/ceramic films. Furthermore, the process has been transformed by the addition of a plasma within the deposition chamber, which provides control of film nucleation and growth kinetics to allow the production of coatings with previously unachievable properties. Variants of this process are given in Fig 1b. The first part of this chapter therefore discusses some basic plasma principles and then outlines some aspects of the influence of plasma bombardment on coating morphology. Then

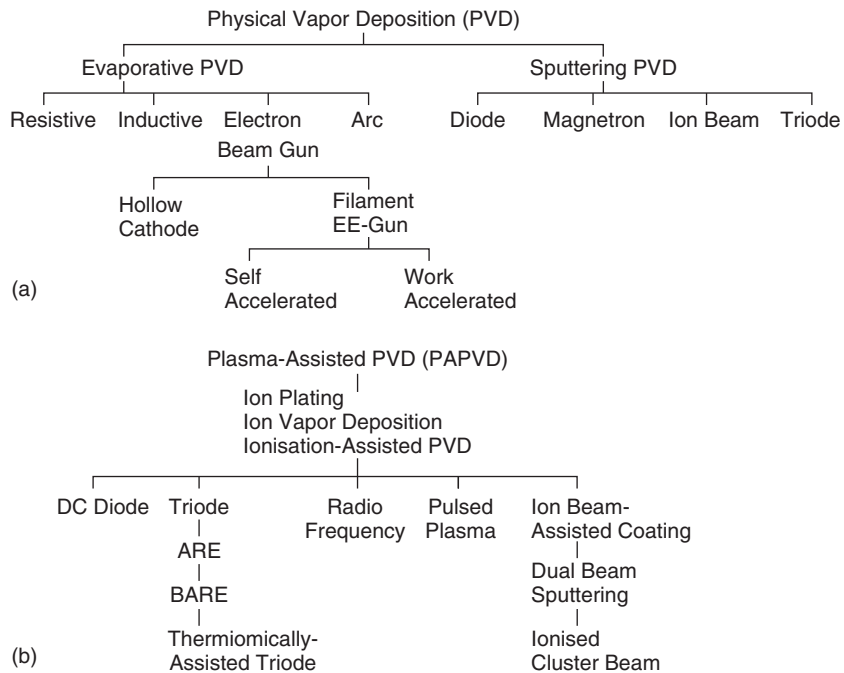


Figure 1 (a) Some physical vapor deposition coating processes; (b) plasma derivatives.

the main evaporative and sputtering processes are discussed followed by a description of some hybrid techniques.

2 GLOW DISCHARGE PLASMA

As mentioned above, plasma assistance is now used routinely in PVD processes. Usually the substrate is the cathode in such systems and is bombarded by ions prior to and during film growth. To explain the mechanisms occurring, it is convenient to consider a simple dc diode argon plasma, shown schematically in Fig. 2.³ The “negative glow” is the visible plasma, which is a partially ionized gas with equal numbers of positive and negative charges (ions and electrons). The plasma is thus virtually field free; most of the voltage in the discharge is dropped across the cathode sheath, which is also called the cathode fall region. Positive ions are thus accelerated toward the cathode. However, they may undergo charge exchange collisions with neutral atoms as they traverse the sheath region. This means that both accelerated neutrals and ions will arrive at the cathode with a range of energies. Davis and Vanderslice⁴ showed how the energy spectra can be calculated for a dc diode discharge, and Fancey and Matthews^{3,5} have indicated how these results will apply in plasma-assisted (PA) PVD systems. A key parameter is the ratio between L (the cathode fall distance) and λ (the mean free path for charge exchange). As we shall see, modern PAPVD systems utilize ionization enhancement devices, which reduce L and thus L/λ , which results in few collisions in the sheath and most of the ions arriving with the full acceleration energy. This gives the benefit that more will thus have the energy required to create the surface nucleation and growth effects necessary for optimized coatings. Some of these energy requirements are listed in Table 1.

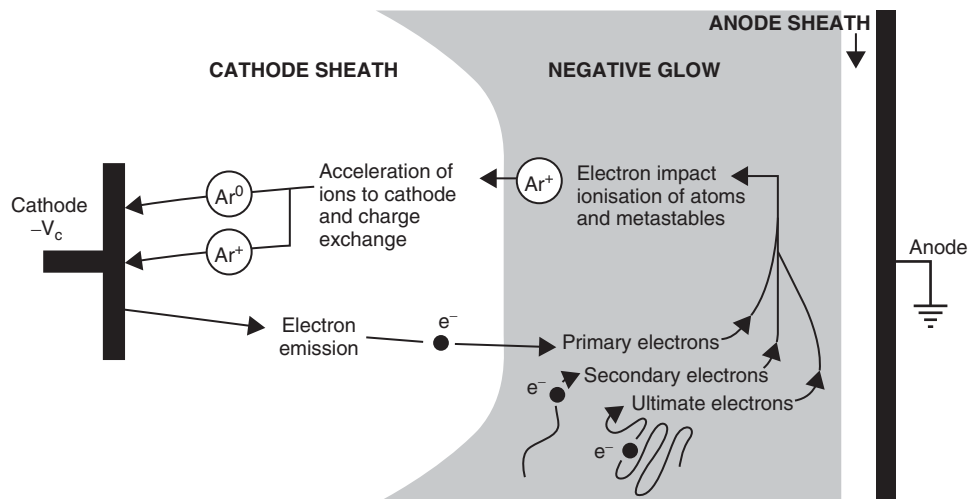


Figure 2 Schematic of argon dc diode glow discharge.

The original PAPVD system was patented by Berghaus⁶ in 1938, but it was not until Mattox⁷ coined the term *ion plating* in the early 1960s that the potential of the process was recognized. These researchers used the dc diode configuration. Enhancing this system means not only that the ions will not undergo energy-reducing collisions but also that there will be more ions available. Several measures can be used to define the actual level of ionization. For example, the term “ionization efficiency” has been used to define the percentage ratio between the ions arriving at the substrate surface during deposition and the total bombardment.^{3,5,8,9} Usually the latter parameter is derived from the background chamber pressure. This, however, assumes that the arrival rate of depositing species is very much lower than the arrival rate of “background” gas atoms. Others have taken a rather different view and have chosen to use the ratio between the ion current and the metal species arrival rate as the determining parameter.^{10–12} This latter approach was first used by researchers who modeled growth using computer routines to predict the coating structure.^{13,14} Both of these approaches can be criticized to some extent. The former becomes increasingly invalid for higher deposition rates and lower chamber pressures, while the latter does not take into account the background chamber pressure (which is known to have a very dominant influence on coating structure). Notwithstanding these deficiencies, both of these measures of process effectiveness confirm that the most important single goal in plasma-assisted PVD is to achieve an adequate level of ionization, and different processes ensure this occurs by a variety of means, as discussed in the following sections.

3 FILM FORMATION AND GROWTH

As Jehn¹⁵ has pointed out, a number of film nucleation and growth models have been published. However, these typically apply only for certain idealized conditions, such as a single-phase substrate surface, no alloy or compound formation, and deposition under ultrahigh-vacuum conditions. Such models are not strictly applicable for ceramic deposition, especially under plasma bombardment conditions. It is thus perhaps more instructive to consider here those models that rely on empirical observations to describe the nature of film growth morphology under different conditions.

Table 1 Important Mechanisms in Plasma-Assisted PVD

Mechanism	Description	Effects	Energy Requirements
Ionization	Electrons emitted from substrate accelerate across substrate (cathode) sheath to gain sufficient energy for electron impact ionization in negative glow region	Maintains discharge in diode configuration; provides means of ionization in addition to any enhancement applied to process	For argon and many other gases (e.g., nitrogen), maximum collision cross section for electron impact ionization typically 70–100 eV
Substrate surface contamination	Desorption of adsorbed impurities on the substrate surface prior to deposition	Prevents, for example, contaminants reducing adhesion between coating and substrate	Several eV
Atom mobility	Removal/surface diffusion of adsorbed atoms on the growing coating surface	Promotes coating densification	Several eV
Atomic displacement	Atoms in the substrate and coating displaced from their normal sites, creating lattice defects	Can lead to, for example, intermixing of substrate and coating atoms	Thresholds in the region of 20–50 eV
Sputtering	Substrate and subsequent growing coating is sputtered	Increased defect densities promote rapid interdiffusion	Thresholds (argon bombarding metals) in the region of 15–35 eV
		Improves coating–substrate adhesion	
Entrapments	Support gas (argon) incorporated in coating during deposition	May also increase coating densification	Argon entrapment probability believed to be very low, below 100 eV
		Can be used to clean substrate prior to deposition	
		Promotes atomic mixing, which can improve coating–substrate adhesion and coating densification	
		Gas atoms may cluster within coating to form bubbles	
		May be detrimental to coating properties	

Source: From Ref. 3.

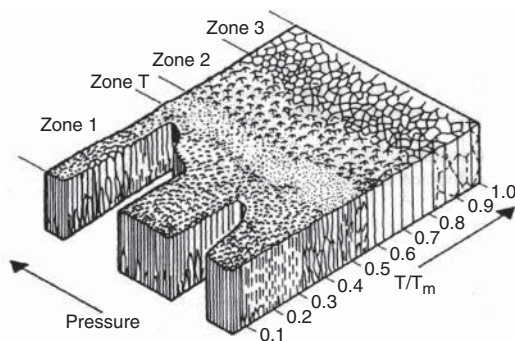


Figure 3 Thornton zone diagram: T (K) is substrate temperature and T_m (K) is melting point of coating material.

The most widely reported of these models is the Thornton structure zone diagram¹⁶ (Fig. 3), which was an extension on the earlier zone model of Movchan and Demchishin.¹⁷ The zone models are based on the observation that PVD coatings deposited under high-pressure (i.e., medium-vacuum) conditions tend to be porous and columnar. This is known as a zone 1 structure. At higher temperatures and/or lower pressures the coatings become more fine grained, densely packed, and fibrous. Thornton called this a zone T structure. At higher temperatures the films exhibit a dense columnar morphology known as zone 2. At even higher temperatures the films exhibit an equiaxed grain structure (zone 3) similar to a recrystallized solid. Under nonbombardment conditions zone 2 morphologies would normally occur only at substrate temperatures above about 50% of the melting temperature of the coating material (i.e., over 1200 K for many ceramics). The remarkable effect of ion bombardment is that it can induce such structures at substrate temperatures of less than one-tenth of the coating material melting temperature.

4 PROCESS DETAILS

4.1 Evaporative Processes

Although there are several ways by which the metal can be evaporated in PAPVD, as given in Fig. 1, the electron beam and arc techniques dominate. These are considered separately here.

In electron beam evaporation the electron source is most often a heated filament^{8,9,18,19}; alternatively, the “hollow-cathode” effect can be utilized to produce a collimated or diffuse beam of electrons from a confining tube^{20–23} (Fig. 4). In either case the resultant beam may be magnetically focused onto a crucible to achieve evaporation, and this crucible may itself be biased positively to permit greater control of the power input (and also the plasma conditions in the deposition chamber)^{24–26} (Fig. 5). Many systems that utilize electron beam evaporation adopt configurations in which the electron energy is relatively low (e.g., tens of electron volts) with a high current flow. Additionally, the beam is frequently directed through the deposition chamber volume. Such arrangements ensure that the electron beam significantly enhances the ionization. In some cases the electrons may be used to directly preheat the components to be coated by biasing them positively prior to the initiation of the precoating and deposition plasmas.^{24,25} Additional control of the plasma during deposition may be achieved by the incorporation of magnetic confinement or by features such as an additional positive electrode^{27,28} and/or electron emission source^{29,30} (Fig. 6).

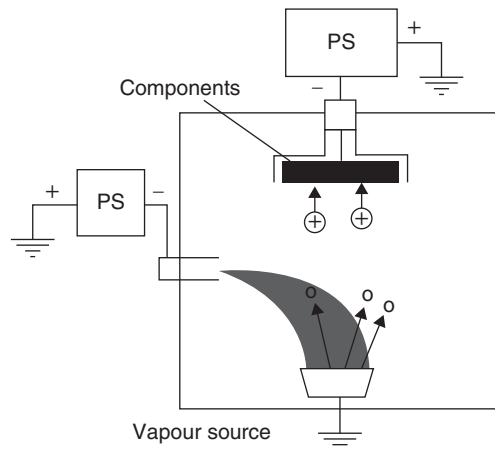


Figure 4 Schematic of hollow-cathode discharge (HCD) electron beam (EB) PVD process.

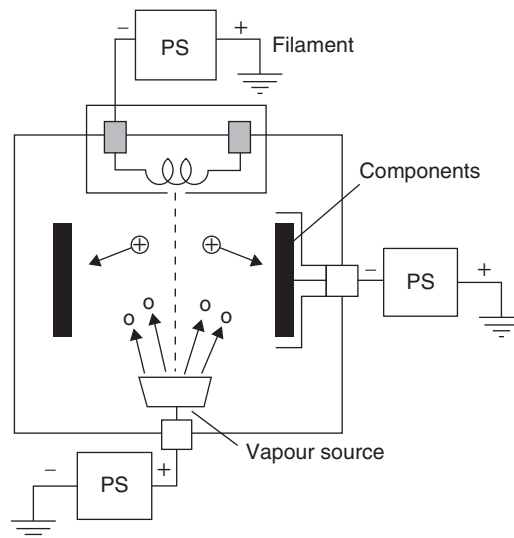


Figure 5 Schematic of low-voltage thermionic EB PVD process.

Such modifications are bound to induce inhomogeneity into the plasma, and this can have a detrimental influence on the uniformity of the resultant coatings.^{31,32} However, the impact of such effects can be minimized by appropriately locating the enhancement device relative to the vapor source.^{3,5}

An advantage of systems that incorporate ionization enhancement that is independent of the vapor source is that more effective plasma heating can be achieved prior to deposition. Also, independent enhancement provides the possibility to achieve controlled plasma diffusion processes (such as ion/plasma nitriding or carburizing),^{33–35} which can be particularly useful as precoating surface treatments. This allows the formation of “duplex” surfaces which combine diffusion treatments and coatings.^{8,36,37}

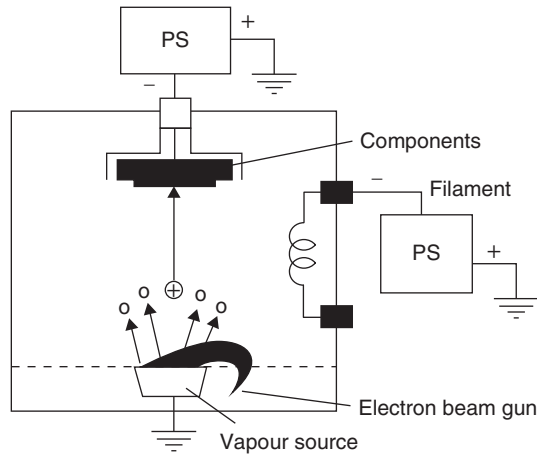


Figure 6 Schematic of thermionically assisted triode PVD process.

An alternative evaporative process is arc evaporation, which has the advantage that it generates copious quantities of energetic electrons and achieves a high degree of metal ionization—both desirable characteristics for the optimization of PAPVD systems.^{38–41} Figures 7 and 8⁴² illustrate the arc evaporation process. Martin⁴² has described this as follows. The emission site of a discrete cathode spot is active for a short period, extinguishes, and then reestablishes close to the original site. The emission site is a source of electrons and atoms of the source material, which are subsequently ionized above the arc site. The ions can either flow back toward the cathode or be accelerated outward toward anodic surfaces, together with

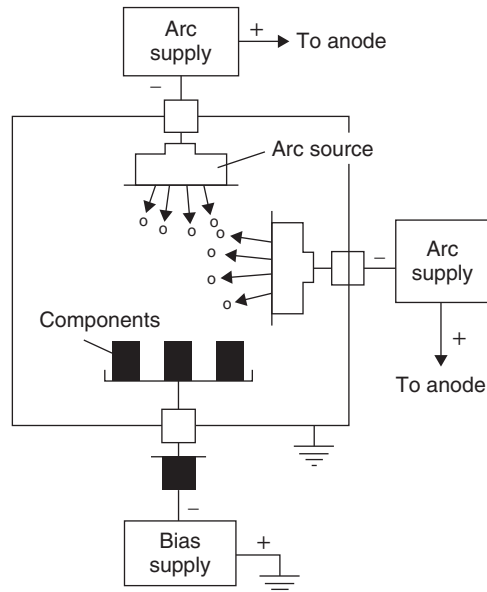


Figure 7 Schematic of arc evaporation PVD process.

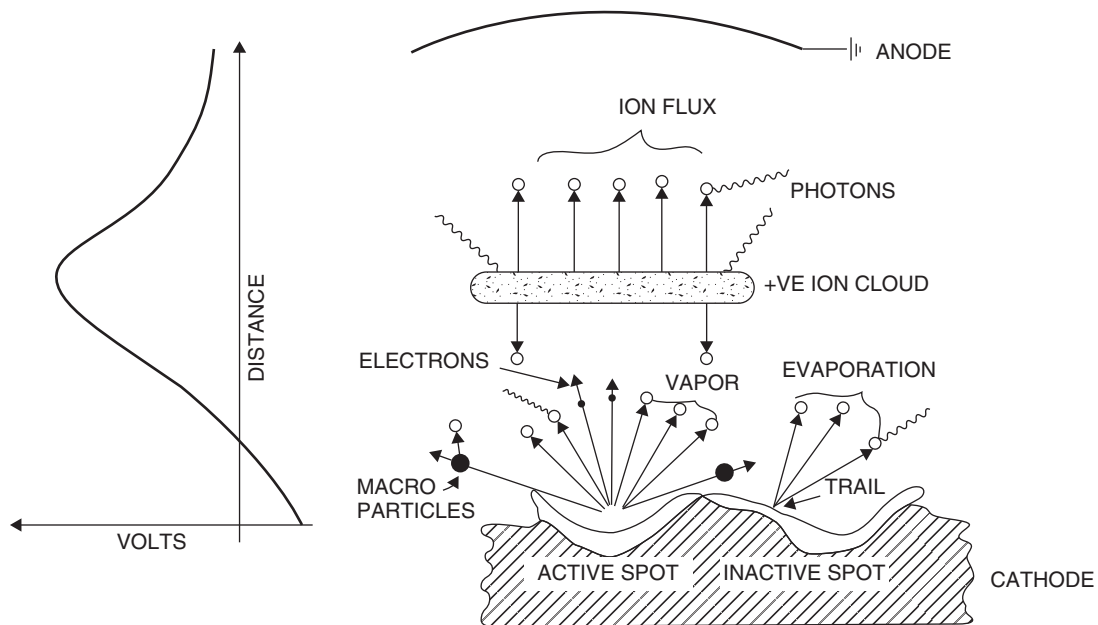


Figure 8 Mechanisms occurring near arc spot and potential distribution.

the electrons. In the case of titanium the degree of ionization of the metal can be over 80% and typical ion energies are 40–100 eV.

Certain designs of arc sources can also be run as magnetron sputtering sources.^{43,44} It has been found that if the arc source is run during the sputter-cleaning phase, then coating adhesion can be improved.^{45–47}

4.2 Sputter Deposition Processes

The simplest sputtering plasma is that of dc diode glow discharge and thus will provide the model in which the basics of the sputtering processes are discussed. In diode sputtering there are only two electrodes, a positive anode and a negatively biased cathode. If the potential applied between these two electrodes is constant over time, it is termed a dc diode discharge. However, if this potential changes with time it may be either a pulsed-diode discharge (unipolar or bipolar), which has come into increased use for sputtering applications, or if the frequency is sufficiently high and reverses in polarity, it is a radio frequency (rf) diode discharge. Each of these has relative advantages and disadvantages, particularly in terms of ceramic coatings. However, all diode plasmas have certain similar mechanisms; in particular, the previous comments with regard to ions traversing the sheath and acquiring sufficient energy to permit kinetically induced mechanisms will apply.

dc Diode Sputtering

Although dc diode sputtering was used in the nineteenth century to deposit thin metallic films on mirrors, it was not until reliable rf sputtering was developed in the 1970s that it was possible to sputter insulating materials. Since many ceramic thin-film materials are good insulators, most early applications of sputtering were thus limited to the deposition of metallic coatings. Therefore dc diode sputtering would not have been widely used in the deposition of ceramic

thin films if the technology necessary for high-rate reactive sputtering from a metallic target had not been developed in the early 1980s.^{48,49}

Provided the target material is sufficiently conductive to avoid the buildup of charge on its surface, the target can be effectively sputtered using simple dc diode geometry. Typically, in dc diode sputtering a potential of about 1000 V is applied between the cathode and anode, and gas pressures in the range from a few millitorr up to 100 mtorr are used to generate a plasma. The substrates are immersed in the discharge, which expands to fill the chamber. The substrates can be heated, cooled, held at ground potential, electrically isolated (i.e., floated), or biased relative to the plasma, each of which affects the properties of the resulting thin films. In general, substrates in a dc diode discharge are subject to a significant amount of electron heating, making this process unsuitable for many semiconductor applications. However, in the case of ceramic thin films for engineering applications this additional heating can be advantageous.

rf Diode Sputtering

The use of an oscillating power source to generate a sputtering plasma offers several advantages over dc methods. The main one is that when the frequency of oscillation is greater than about 50 kHz, it is no longer necessary for both electrodes to be conductive because the electrode can be coupled through an impedance⁵⁰ and will take up a negative dc offset voltage due to the greater mobility of electrons compared to ions in the reversing field. The coupled electrode must be much smaller than the direct electrode to effectively sputter only the insulating (coupled) electrode; this is usually accomplished by utilizing the grounded chamber walls as the other electrode. An impedance matching network is integrated into the circuit between the rf generator and the load to introduce the inductance necessary to form a resonant circuit.

An additional benefit of using rf frequencies above 50 kHz is that the electrons have a longer residence in the negative-glow region and also have sufficient energy to directly ionize the gas atoms; hence, the number of electrons required to sustain the discharge is substantially reduced.^{51,52} This, in turn, means that lower sputtering pressures can be used, reducing the risk of film contamination. The most commonly used frequencies are 13.56 and 27 MHz; these are the frequencies specified by the U.S. Federal Communications Commission (FCC) for medical and industrial use.⁵⁰ The applications of rf sputtering are quite varied and include deposition of metals, metallic alloys, oxides, nitrides, and carbides.^{53–56}

A number of comprehensive reviews and discussions of rf discharges are available in the literature.^{50,57–60} In general, the primary advantages of rf sputtering are:

- Ability to sputter insulators as well as almost any other material
- Operation at lower pressures

Unfortunately, the deposition rates in rf sputtering are often limited by the low thermal conductivity of the insulating target materials, which can lead to the formation of “hot spots” on the target; the hot spots generate stresses that may cause fracture of the brittle target materials. For this reason, it may be preferable to deposit insulating films reactively from a metal source. Although compound materials can be readily sputtered in rf discharges, the resulting films may not be representative of the initial target composition.

Pulsed Power Sputtering

The availability of sophisticated electrical supplies capable of providing controlled pulsed power waveforms with excellent arc control opened up new opportunities in sputter deposition.^{61–63} In particular, medium-frequency (MF) ac and pulse-dc power have proved highly effective in the reactive sputtering of oxide coatings. Scherer et al⁶² reported MF ac magnetron sputter deposition of dielectric Al₂O₃, SiO₂, and Si₃N₄ layers. Several authors have shown that within a frequency range of 10–100 kHz, where ions follow the ac electric field completely,

deposition rates close to dc can be achieved. Combining ac power with feedback-based process stabilization techniques in order to operate in the transition region (i.e., between metallic and fully poisoned target states) resulted in high deposition rates and almost eliminated arcing at the target.

Sproul⁶⁴ described the use of bipolar pulsed power for sputtering of metal targets in oxygen as follows. The polarity of the target power is switched from negative to positive, and during the positive pulse any charging of oxide layers formed on the target surface is discharged when electrons are attracted to the positive surface. During the negative pulse, ions are attracted to the target surface and sputtering takes place initially from all surfaces on the target, even those that have formed a compound, since the charge on that surface has been neutralized during the positive pulse.

Bipolar pulse power can be either symmetric or asymmetric, depending on the relative positive and negative maximum voltages.⁶⁵ Pulsed dc power is said to be symmetric if these voltages are equal. With advances in power supply technologies it proved possible to control the on and off time for the pulses.^{65,66}

Dual-cathode MF ac power has found wide applicability in industrial reactive deposition systems for dielectric coatings, while symmetric bipolar pulsed dc power is less widely used. Two sputter targets mounted side by side are both connected to the same MF ac power supply. One power lead is connected to one target, and the other to the second target. With this system, one of the sputter targets is briefly the anode, while the other is the cathode, and this continuously swaps with reversals in polarity. Sputtering from the cathode surface during the negative pulse keeps the target surface clean, and when it switches to act as an anode it is not covered by an oxide. This procedure avoids the disappearing anode problem, which can occur in pulsed dc sputtering of oxides when all surfaces in the chamber become covered with an insulating oxide.

Asymmetric bipolar pulsed dc has unequal pulse heights and will be applied to one target. Usually, the negative pulse voltage is greater than the positive one. The width of the positive pulse (or “pulse off”) is typically 10–20% of the negative one (“pulse on”); thus a high proportion of the cycle is spent in the sputtering mode, providing deposition rates that are near to those for nonreactive dc sputtering of metals, provided that pulsed dc power is combined with effective closed-loop reactive sputtering process control. For example, Ref. 67 reports reactive sputtering of aluminum in oxygen at 76% of the metal rate under nonreactive conditions. Typical frequencies used are 20–200 kHz. Sproul⁶⁴ states that the frequency selected depends on the material being deposited. Whereas no arcing occurs for titanium dioxide deposition at a pulse frequency of 30 kHz, it takes 50–70 kHz to prevent arcing in aluminum oxide deposition.

A recent development in pulsed power technology in sputtering has been high-power impulse magnetron sputtering (HIPIMS or HPPMS).^{68–70} In this process the power is applied to the sputter target in pulses of low duty cycle (<10%) and frequency (<10 kHz). This allows high instantaneous power levels (>1000 W/cm²) to be applied, which results in high ionization of the sputtered metal. In recent years this technology has gone through significant advances, and peak power levels of several megawatts are now reported. Various advantages are claimed for this technology, including reduced stress levels in films (allowing thicker coating), improved density, and lower porosity.⁷¹ The achievement of wider usage for the HIPIMS technology goes hand in hand with process developments, such as accurate sputter power and/or gas flow control linked to optical emission from the plasma.⁷² This is necessary since (despite earlier ideas to the contrary) HIPIMS can lead to target poisoning in the presence of a reactive gas, which influences the sputter rate.⁷³

Triode Sputtering

Triode sputtering uses a thermionic cathode separate from the sputtering target to sustain the plasma. The target electrode then extracts ions from the plasma. This additional electron source,

typically either a simple biased conductor or a thermionic electron emitter, provides a means of sustaining the discharge that is independent of the secondary electron generation at the cathode. Thus, the discharge may be maintained at pressures as low as 10^{-3} Pa (10^{-5} torr) and at discharge voltages as low as 40 V.⁷⁴ By varying the emission of the electron source, the discharge current can be varied independently of the sputtering voltage, allowing high ion densities at the target and substrate while maintaining a low discharge potential. Triode sputtering, both dc and rf, has been used successfully to deposit films of a great variety of materials for semiconductor, wear-resistant, optical, and other coating applications.^{74–76} The primary advantages of triode sputtering are⁷⁶:

- Lower discharge pressures
- Lower discharge voltages
- Higher deposition rates
- Independent control of the plasma density and the bombardment conditions of the sputtering target

The major weaknesses of triodes are that they:

- Are often more complicated to use
- Can increase film contamination from the electron source
- Are difficult to scale up for industrial processing
- May not be suitable in temperature-sensitive and reactive processes because of the electron source

Magnetron Sputtering

Magnetron sputtering has already been mentioned in regard to pulsed dc power. It is a variant on the diode sputtering process, in which a magnetic field is used to trap electrons in the vicinity of the target and to spiral them round a “racetrack,” thereby increasing the degree of ionization occurring and therefore the sputtering rate. Probably the greatest research advances in PAPVD in recent years have occurred around the magnetron sputtering process. The benefits of magnetron sputtering have been known for several decades (i.e., in terms of increasing the deposition rate). However, initially sputtering was found to be less effective as a means of producing hard ceramic coatings on tools and components than electron beam and arc evaporation. This was because the levels of ionization achieved were considerably less than those possible with the enhanced plasmas which are characteristic of these other methods.

The most important step in improving the competitiveness of magnetron sputtering was reported by Window and Savvides^{77,78} in 1986, when they described their studies on the unbalanced magnetron (UBM) effect. They found that the ion flux to the substrate could be considerably increased by running magnetron cathodes in the “unbalanced” mode (UM). They identified three main magnetic arrangements (Fig. 9). In the type I configuration all the field lines originate from the central magnet, with some not passing into the outer magnet. In the intermediate case all of the field lines starting on the central pole go to the outer pole. In the type II configuration all of the field lines originate in the outer magnet, with some not passing to the central pole. Following the work of Window and Savvides, a number of researchers began building upon these concepts.⁷⁶ In the late 1980s Sproul and his co-workers researched multicathode high-rate reactive sputtering systems.^{79,80} They studied different magnetic configurations, principally in the dual-cathode arrangement shown schematically in Fig. 10. They found that by strengthening the outer magnets of a magnetron cathode with NdFeB and arranging two of these magnetrons in an opposed closed-field configuration (i.e., with opposite poles facing each other), the substrate bias current could be increased to well over 5 mA/cm^2 at a

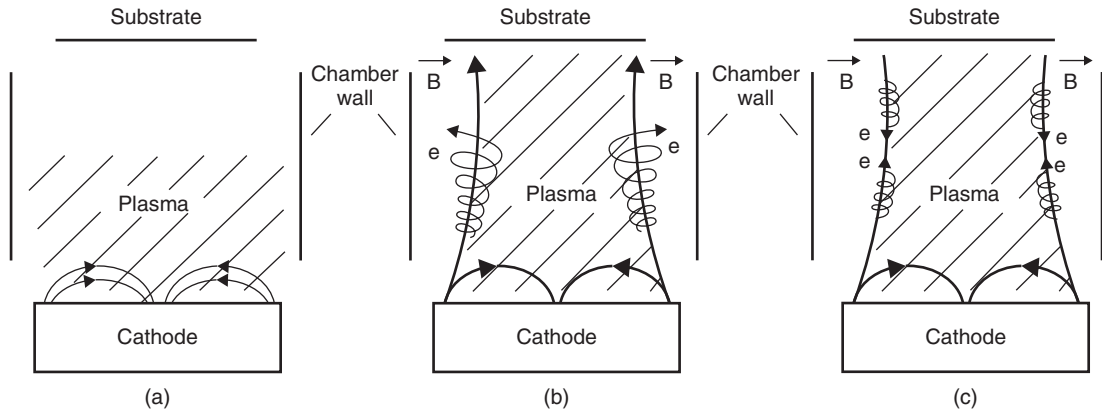


Figure 9 Three main magnetron configurations.

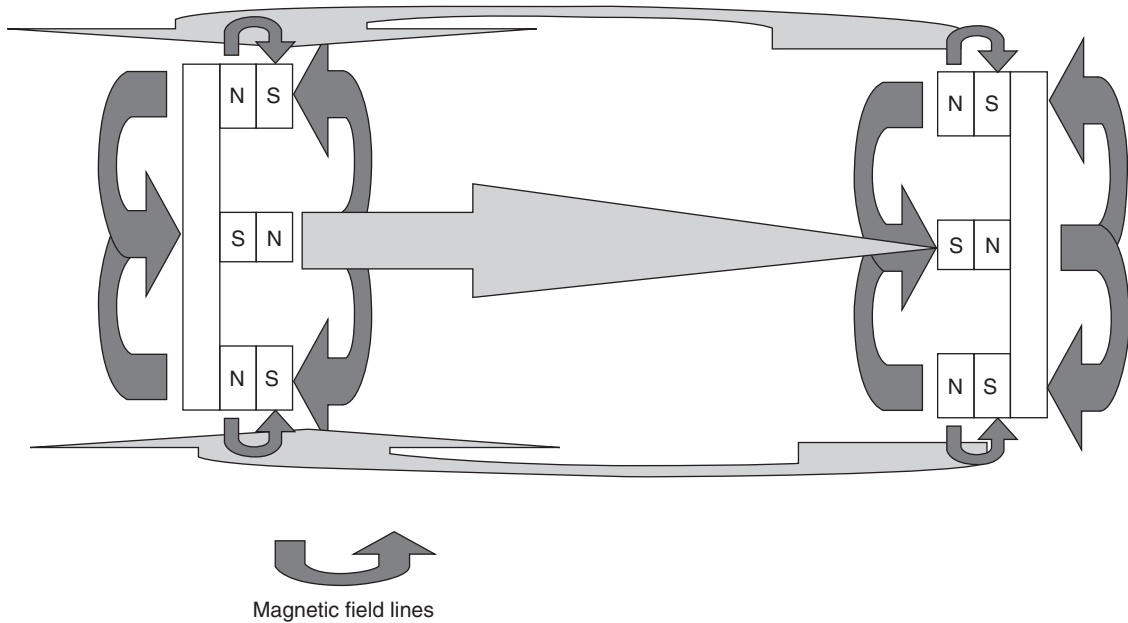


Figure 10 Schematic of opposed unbalanced magnetron configuration.

pressure of 5 mtorr. Under these conditions they were able to deposit hard well-adhered TiN coatings. Another person utilizing the opposed magnetron configuration at that time was Tomi-naga.⁸¹ He presented an arrangement whereby the permanent magnetic field could be increased by external electromagnetic coils. He stated that this allowed the trapping of electrons by the closed-field arrangement to increase ionization or to allow lower working gas pressure with the same level of ionization as achievable with conventional magnetron sputtering.

Another group that has researched confined and unbalanced magnetron sputtering is that led by Musil and Kadlec,^{82,83} who in early collaboration with Munz investigated a circular planer unbalanced magnetron arrangement surrounded by two magnetic coils and a

set of permanent magnets. They describe this as a multipolar magnetic plasma confinement (MMPC) and cite high ion currents at substrates, even when located at large distances from the magnetron.

Howson and co-workers have also studied unbalanced magnetrons, especially for high-power, large-area applications,^{84,85} and thus demonstrated how additional anodes or electromagnets can be used to control the ion current to the substrate. This may be seen as an extension of early work by Morrison,⁸⁶ who demonstrated that a “magnetically hidden” anode placed in a magnetron sputtering system could double the plasma density and provide a more uniform plasma throughout the chamber.⁸⁷ Hofman et al. have described an improved magnetron sputtering system which combines the magnetic anode effect and the opposed closed-field unbalanced magnetron.⁸⁸ Leyendecker et al. utilize a further modification on this theme.⁸⁹

An innovation to the magnetron sputtering technique was described by Rossnagel and Hopwood.⁹⁰ They placed a rf coil between the target and the substrate to increase the degree of ionization of the coating species. In the case of aluminum he achieved ionization levels of up to 80%. Although the technique was developed for metal films for use in metallization of small-aspect-ratio semiconductor structures, it has been shown that reactive deposition of ceramics can be made more effectively by this method. In particular the use of this approach in the reactive deposition of crystalline alumina has been achieved at much lower temperatures than previously reported.^{91,92}

Currently, one of the most significant developments in magnetron sputtering technology for large-scale production use, which gives much improved utilization of the sputter target material and longer target lifetimes, is the rotatable magnetron. This system is now widely used for large-area glass and web coating.⁹³ A key development was the dual ac rotatable system,⁹⁴ which is key for reactive deposition of insulating films.

Another important development, for planar magnetron geometries, is the use of moving magnets. These result in improved target erosion profiles (with target material utilization increased considerably).⁹⁵ Such systems are important for semiconductor processing.

4.3 Beam Processes

A third group of processes that can be included under the PVD heading are those that utilize ion or laser beams, either to influence the growing film or to produce the coating flux. Dearnaley⁹⁶ has cited the four ion beam processes, shown in Fig. 11 as representing the effects that ion beams can have on surfaces. The ion-assisted coating (IAC) or ion-beam-assisted deposition (IBAD) methods compare most closely with the plasma-based PVD processes described earlier.⁹⁷ Hubler and Hirvonen⁹⁸ cite the two geometries shown in Fig. 12 as being the most common of these. According to these authors, what distinguishes IBAD from the other PVD methods is that the source of vapor and the source of energetic ions are separated into two distinct pieces of hardware. In many plasma-based PVD processes the evaporant flux and the ion flux are derived by extraction from the plasma, whereas in IBAD they can be controlled independently. The other major difference between the plasma techniques and IBAD is the chamber pressure used. Typically PAPVD processes are carried out at pressures of several millitorrs, whereas IBAD techniques must usually operate in a near-collision-free pressure regime, better than 10^{-5} torr. This means that IBAD is often restricted to line-of-sight applications. Ceramics such as ZrO₂,⁹⁹ TiO₂,¹⁰⁰ Al₂O₃,¹⁰¹ Si₃N₄,¹⁰² AlN,¹⁰³ TiN,¹⁰⁴ BN,¹⁰⁵ and TiC¹⁰⁶ have been deposited by IBAD methods.

There is also considerable interest in processes to deposit diamondlike carbon (DLC) films using IBAD methods or even by direct gas deposition from an ion or plasma source.^{107–109} A further variation on this theme is the filtered-arc source^{110–112} that utilizes a deflection coil to eliminate macrodroplets and produce a beam of ionized atoms, which can be carbon (for DLC films) or a metal such as titanium for reactive deposition of a ceramic. Another important beam

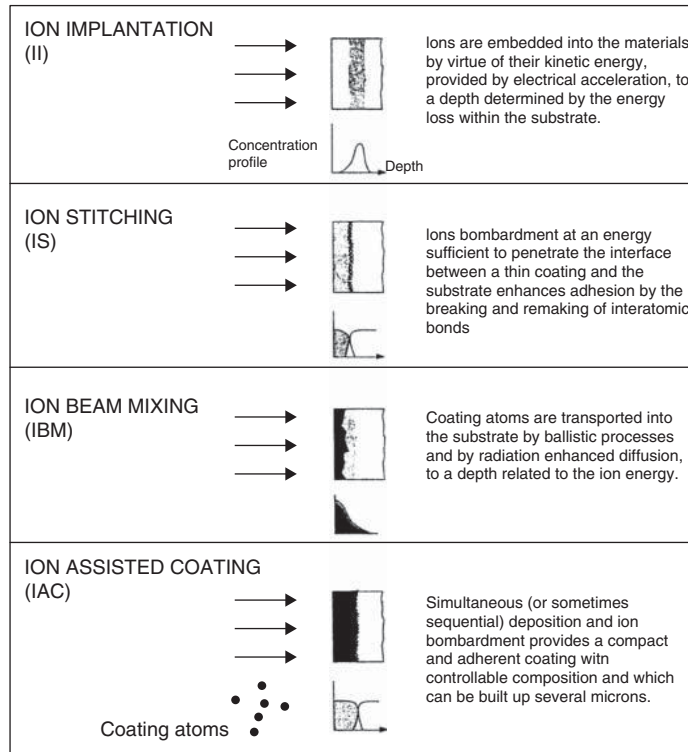


Figure 11 Four main ion-beam-based surface treatment techniques.

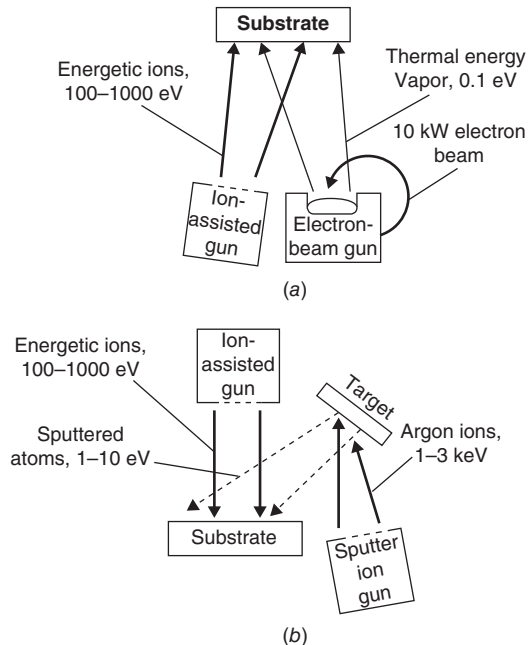


Figure 12 Two common ion beam PVD techniques: (a) ion-beam-assisted deposition (IBAD); (b) dual-ion-beam sputtering (DIBS).

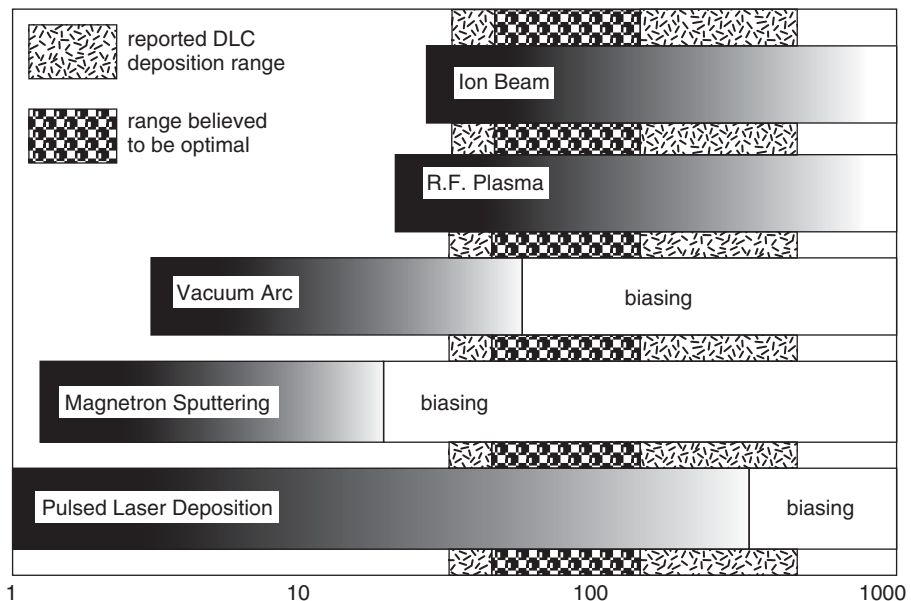


Figure 13 Approximate energy ranges of deposited particles produced by selected PVD techniques. Energy regions corresponding to DLC formation are also shown.

technique in PVD is pulsed laser deposition. According to Voevodin and Donley¹¹³ this is especially appropriate for DLC coatings, as the controlled energy range achievable for the arriving species is wider than other techniques (Fig. 13). Also flux control and species control are said to be excellent; for example, nonhydrogenated DLC can be produced. Several ceramics have also been deposited by this method.^{114,115}

5 FINAL COMMENTS

Plasma or ion assistance in PVD processes is making possible the creation of coating materials and structures which are unachievable by other deposition techniques. Given this advantage and also the fact that these processes are environmentally friendly, we can expect to see their increasing use in many engineering applications. This is especially the case since the latest theories on friction and wear are pointing toward nanocomposite coatings of finely controlled compositions, giving enhanced mechanical properties.^{116–118} In addition, there is an increase in the use of the “duplex” combined treatments and coatings mentioned earlier, and these benefit from the latest plasma enhancement and control systems to impart excellent tribological and load-bearing properties even on lightweight alloys such as those based on titanium.¹¹⁹

REFERENCES

1. W. R. Grove, “On the Electro-Chemical Polarity of Gases,” *Philos. Trans. R. Soc. London A*, **142**, 87, 1852.
2. M. Faraday, “The Bakerian Lecture: Experimental Relations of Gold (and other Metals) to Light,” *Philos. Trans.*, **147**, 145, 1857.
3. K. S. Fancey and A. Matthews, in D. S. Rickerby and A. Matthews (Eds.), *Advanced Surface Coatings*, Blackie, London, 1991.

4. W. D. Davis and T. A. Vanderslice, "Ion Energies at the Cathode of a Glow Discharge," *Phys. Rev.*, **131**, 219, 1963.
5. K. S. Fancey and A. Matthews, "Evaporative Ion Plating: Process Mechanisms and Optimization," *IEEE Trans. Plasma Sci.*, **18**, 869, 1990.
6. B. Berghaus, UK Patent No. 510993, 1938.
7. D. M. Mattox, "Film Deposition Using Accelerated Ions," *Electrochem. Technol.*, **2**, 295, 1964.
8. A. Matthews, "Developments in Ionization Assisted Processes," *J. Vac. Sci. Technol.*, **A3**(6), 2354, 1985.
9. A. Matthews, in W. D. Sproul, J. E. Greene, and J. A. Thornton (Eds.), *Physics and Chemistry of Protective Coatings*, AIP Proceedings 149, 1986, American Institute of Physics.
10. I. Petrov, F. Adibi, J. E. Green, L. Hultman, and J.-E. Sundgren, "Average Energy Deposited per Atom: A Universal Parameter for Describing Ion-Assisted Film Growth," *Appl. Phys. Lett.*, **63**, 36, 1993.
11. G. Hakansson, L. Hultman, J.-E. Sundgren, J. E. Greene, and W.-D. Munz, "Microstructures of TiN Films Grown by Various Physical Vapor-Deposition Techniques," *Surf. Coat. Technol.*, **48**, 51, 1991.
12. F. Abidi, I. Petrov, J. E. Greene, L. Hultman, and J.-E. Sundgren, "Effects of High-Flux Low-Energy (20-100eV) Ion Irradiation during Deposition on Microstructure and Preferred Orientation of $Ti_{0.5}Al_{0.5}N$ Alloys Grown by Ultra-High Vacuum Reactive Magnetron Sputtering," *J. Appl. Phys.*, **73**, 8580, 1993.
13. K.-H. Muller, "Ion-Beam Induced Epitaxial Vapor-Phase Growth: A Molecular Dynamics Study," *Phys. Rev. B*, **35**, 7906, 1987.
14. K.-H. Muller, "Molecular Dynamics and Collision Cascade Studies of Ion-Assisted Thin Film Deposition," *J. Vac. Sci. Technol.*, **A5**, 2161, 1987.
15. H. A. Jehn, in W. Gissler and H. A. Jehn (Eds.), *Advanced Techniques for Surface Engineering*, Kluwer Academic, New York, 1992.
16. J. A. Thornton, "Influence of Apparatus Geometry and Deposition Conditions on the Structure and Topography of Thick Sputtered Coatings," *J. Vac. Sci. Technol.*, **11**, 666, 1974.
17. B. A. Movchan and A. V. Demchishin, "Study of the Structure and Properties of Thick Vacuum Condensates of Nickel, Titanium, Tungsten, Aluminium Oxide and Zirconium Dioxide," *Fiz. Metal. Metalloved.*, **28**, 653, 1969.
18. Y. Enomoto and K. Matsubara, "Structure and Mechanical Properties of Ion-Plated Thick Films," *J. Vac. Sci. Technol.*, **12**, 827, 1975.
19. H. K. Pulker, *Coatings on Glass*, 2nd ed., Elsevier, Amsterdam, 1999.
20. J. R. Morley and H. R. Smith, "High Rate Ion Production for Vacuum Deposition," *J. Vac. Sci. Technol.*, **9**, 1377, 1972.
21. S. Komiya and K. Tsuruska, "Physical Vapor Deposition of Thick Cr and Its Carbide and Nitride Films by Hollow-Cathode Discharge," *J. Vac. Sci. Technol.*, **13**, 520, 1976.
22. C. T. Wan, D. L. Chambers, and D. C. Carmichael, "Effect of Processing Conditions on Characteristics of Coatings Vacuum Deposited by Ion Plating," in *Proceedings of the Fourth International Conference on Vacuum Metallurgy*, Iron and Steel Institute of Japan, Tokyo, Japan, 1974, p. 231.
23. S. Komiya and K. Tsuruoka, "Thermal Input to Substrate during Deposition by Hollow-Cathode Discharge," *J. Vac. Sci. Technol.*, **12**, 589, 1975.
24. E. Moll, in W. Gissler and H. Jehn (Eds.), *Advanced Techniques for Surface Engineering*, Kluwer, Dordrecht, 1992.
25. E. Moll and H. Daxinger, U.S. Patent No. 4,197,175, 1980.
26. R. Buhl, H. K. Pulker, and E. Moll, "TiN Coatings on Steel," *Thin Solid Films*, **80**, 265-270, 1981.
27. M. Koboyashi and Y. Doi, "TiN and TiC Coating on Cemented Carbides by Ion Plating," *Thin Solid Films*, **54**, 17, 1978.
28. R. F. Bunshah, "The Activated Reactive Evaporation Process: Developments and Applications," *Thin Solid Films*, **80**, 255, 1981.
29. A. Matthews and D. G. Teer, "Deposition of Ti-N Compounds by Thermionically Assisted Triode Reactive Ion Plating," *Thin Solid Films*, **72**, 541, 1980.
30. G. A. Baum, Report No. RFP-686 UC-25, Dow Chemical Co., Golden, CO, 1967.

31. K. S. Fancey, M. Williams, A. Leyland, and A. Matthews, "The Influence of Process System Characteristics on the Uniformity of Ion Plated Titanium Nitride Coatings," *Vacuum*, **43**, 235, 1992.
32. K. S. Fancey and A. Matthews, "An Investigation into the Variation in Bombardment Intensity from Ion Plating Discharges by Sputter Weight Loss Experiments," *Thin Solid Films*, **193–194**, 171, 1990.
33. A. Leyland, K. S. Fancey, and A. Matthews, "Plasma Nitriding in a Low Pressure Triode Discharge to Provide Improvements in Adhesion and Load Support for Wear Resistant Coatings," *Surf. Eng.*, **7**, 207, 1991.
34. P. R. Stevenson, A. Leyland, M. A. Parkin, and A. Matthews, "The Effect of Process Parameters on the Plasma Carbon Diffusion Treatment of Stainless Steels at Low Pressure," *Surf. Coat. Technol.*, **63**, 135–143, 1994.
35. A. Leyland, D. B. Lewis, P. R. Stevenson, and A. Matthews, "Low Temperature Plasma Diffusion Treatment of Stainless Steels for Improved Wear Resistance," *Surf. Coat. Technol.*, **62**, 608–617, 1993.
36. D. B. Lewis, A. Leyland, P. R. Stevenson, J. Cawley, and A. Matthews, "Metallurgical Study of Low-Temperature Plasma Carbon Diffusion Treatments for Stainless Steels," *Surf. Coat. Technol.*, **60**, 416–423, 1993.
37. A. Matthews and A. Leyland, "Hybrid Techniques in Surface Engineering," *Surf. Coat. Technol.*, **71**, 88–92, 1995.
38. P. A. Lindfors, W. M. Mularie, and G. K. Wehner, "Cathodic Arc Deposition Technology," *Surf. Coat. Technol.*, **29**, 275, 1986.
39. R. L. Boxman and S. Goldsmith, "Cathode-Spot Arc Coatings: Physics, Deposition and Heating Rates, and Some Examples," *Surf. Coat. Technol.*, **33**, 153, 1987.
40. P. J. Martin, D. R. McKenzie, P. P. Netterfield, P. Swift, S. W. Filipczuk, K. J. Muller, C. G. Pacey, and B. James, "Characteristics of Titanium Arc Evaporation Processes," *Thin Solid Films*, **153**, 91, 1987.
41. D. M. Sanders, D. B. Boercker, and S. Falabella, "Coating Technology Based on the Vacuum Arc—a Review," *IEEE Trans. Plasma Sci.*, **18**, 883, 1990.
42. P. J. Martin, "Cathodic Arc Deposition," in D. A. Gleeker and S. I. Shah (Eds.), *Handbook of Thin Film Process Technology*, Institute of Physics Publishing, Bristol, U.K., 1995.
43. P. A. Robinson and A. Matthews, "Characteristics of a Dual Purpose Cathodic Arc/Magnetron Sputtering System," *Surf. Coat. Technol.*, **43–44**, 288, 1990.
44. Hauzer Holding BV, European Patent Application PCT/EP90 01032, 1990.
45. W.-D. Munz, F. J. M. Hauzer, D. Schulze, and B. Buil, "A New Concept for Physical Vapor Deposition Coating Combining the Methods of Arc Evaporation and Unbalanced Magnetron Sputtering," *Surf. Coat. Technol.*, **49**, 161, 1991.
46. W.-D. Munz, D. Schulze, and F. J. M. Hauzer, "A New Method for Hard Coatings: ABS™ (Arc Bond Sputtering)," *Surf. Coat. Technol.*, **50**, 169, 1992.
47. W.-D. Munz, K. Vannisselroy, R. Tietma, T. Turkmans, and G. Keiren, "An All-Round Performer in the Physical Vapor Deposition Laboratory," *Surf. Coat. Technol.*, **58**, 205, 1993.
48. W. D. Sproul, "Very High Rate Reactive Sputtering of TiN, ZrN and HfN," *Thin Solid Films*, **107**, 141–147, 1983.
49. W. D. Sproul and J. A. Tomashek, U.S. Patent No. 4,428,811, 1984.
50. R. J. Hill (Ed.), *Physical Vapor Deposition*, Temescal, Livermore, CA, 1986.
51. J. L. Vossen and J. J. Cuomo, "Glow Discharge Sputter Deposition," in J. L. Vossen and W. Kern (Eds.), *Thin Film Processes*, Academic, New York, 1978, pp. 12–73.
52. J. A. Thornton, "Coating Deposition by Sputtering," in R. F. Bunshah (Ed.), *Deposition Technologies for Thin Films and Coatings*, Noyes, Park Ridge, NJ, 1982, pp. 170–243.
53. B. Chapman, *Sputtering, Glow Discharge Process*, Wiley, New York, 1980, pp. 177–296.
54. J.-E. Sundgren, B.-O. Johansson, and S. E. Karlsson, "Mechanism of Reactive Sputtering of Titanium Nitride and Titanium Carbide I: Influence of Plasma Process Parameters on Film Composition," *Thin Solid Films*, **105**, 353–366, 1983.
55. J.-E. Sundgren, B.-O. Johansson, S.-E. Karlsson, and H. T. G. Hentzell, "Mechanism of Reactive Sputtering of Titanium Nitride and Titanium Carbide II: Morphology and Structure," *Thin Solid Films*, **105**, 367–384, 1983.

56. J.-E. Sundgren, B.-O. Johansson, S.-E. Karlsson, and H. T. G. Hentzell, "Mechanism of Reactive Sputtering of Titanium Nitride and Titanium Carbide III: Influence of Substrate Bias on Composition and Structure," *Thin Solid Films*, **105**, 385–393, 1983.
57. J. L. Cecchi, "Introduction to Plasma Concepts and Discharge Configurations," in S. M. Rossnagel, J. J. Cuomo, and W. D. Westwood (Eds.), *Handbook of Plasma Processing Technology*, Noyes, Park Ridge, NJ, 1990, pp. 14–69.
58. J. S. Logan, "RF Diode Sputter Etching and Deposition," in S. M. Rossnagel, J. J. Cuomo, and W. D. Westwood (Eds.), *Handbook of Plasma Processing*, Noyes, Park Ridge, NJ, 1990, pp. 140–159.
59. B. Chapman, "RF Discharges," in *Glow Discharge Processes*, Wiley, New York, 1980, pp. 139–175.
60. G. N. Jackson, "RF Sputtering," *Thin Solid Films*, **5**, 209–246, 1970.
61. S. Schiller, K. Goedicke, V. Kirchoff, and T. Kopte, in *Proceedings of the 38th Annual Technical Conference of the Society of Vacuum Coaters*, Chicago, April 1995, SVC, Albuquerque, NM, 1995.
62. M. Scherer, J. Schmitt, R. Latz, and M. Schanz, J. Vac. "Reactive Alternating Current Magnetron Sputtering of Dielectric Layers." *Sci. Technol., A*, **10**, 1772, 1992.
63. S. Schiller, K. Goedicke, J. Reschke, V. Kirchhoff, S. Schneider, and F. Milde, "Pulsed Magnetron Sputter Technology," *Surf. Coat. Technol.*, **61**, 331–227, 1993.
64. W. D. Sproul, "Advances in Reactive Sputtering," in *Proceedings of the 39th Annual Technical Conference of the Society of Vacuum Coaters*, Philadelphia, May 1996, SVC, Albuquerque, NM, 1996.
65. R. A. Scholl, "Reactive PV Deposition of Insulators," in *Proceedings of the 39th Annual Technical Conference of the Society of Vacuum Coaters*, Philadelphia, May 1996, SVC, Albuquerque, NM, 1996.
66. J. C. Sellers, "Asymmetric Bipolar Pulse DC—An Enabling Technology for Reactive PVD," in *Proceedings of the 39th Annual Technical Conference of the Society of Vacuum Coaters*, Philadelphia, May 1996, SVC, Albuquerque, NM, 1996.
67. J. M. Schneider, A. A. Voevodin, M. S. Wong, W. D. Sproul, and A. Matthews, "Very-High-Rate Reactive Sputtering of Alumina Hard Coatings," *Surf. Coat. Technol.*, **96**, 262, 1997.
68. J. Bohlmark, M. Lattemann, J. T. Gudmundsson, A. P. Ehiasarian, Y. Aranda Gonzalvo, N. Brenning, and U. Helmersson, "The Ion Energy Distributions and Ion Flux Composition from a High Power Impulse Magnetron Sputtering Discharge," *Thin Solid Films*, **515**, 1522, 2006.
69. U. Helmersson, M. Lattemann, J. Bohlmark, A. P. Ehiasarian, and J. T. Gudmundsson, "Ionized Physical Vapor Deposition (IPVD): A Review of Technology and Applications," *Thin Solid Films*, **513**, 1, 2006.
70. K. Sarakinos, J. Alami, and S. Konstantinidis, "High Power Pulsed Magnetron Sputtering: A Review on Scientific and Engineering State of the Art," *Surf. Coat. Technol.*, **204**, 1661, 2010.
71. W. D. Sproul, "An Exciting Time to Be a Sputterer," *SVC Bull.*, Fall, 22, 2008.
72. M. Audronis, V. Bellido-gonzalez, and B. Daniel, "Control of Reactive High Power Impulse Magnetron Sputtering Processes," *Surf. Coat. Technol.*, **204**, 2159, 2010.
73. M. Audronis, G. Abrasonis, F. Munnik, R. Heller, P. Chapon, and V. Bellido-Gonzalez, "Diffusive Racetrack Oxidation in a Ti Sputter Target by Reactive High Power Impulse Magnetron Sputtering," *J. Phys. D*, **45**, 375203, 2012.
74. S. L. Rohde, S. A. Barnett, and C.-H. Choi, J. Vac. "An Ultrahigh Vacuum, Low-Energy Ion-Assisted Deposition System for III-V Semiconductor Film Growth," *Sci. Technol.*, **A7**(3), 2273–2279, 1989.
75. G. Mah, C. W. Nordin, and V. F. Fuller, J. Vac. "Structure and Properties of Sputtered Titanium Carbide and Titanium Nitride Coatings," *Sci. Technol.*, **11**(1), 371–373, 1974.
76. S. L. Rohde, in *ASM Handbook*, Vol. 5: *Surface Engineering*, ASM International, Metals Park, OH, 1994.
77. B. Window and N. Savvides, "Charged Particle Fluxes from Planar Magnetron Sputtering Sources," *J. Vac. Sci. Technol.*, **A4**(2), 196, 1986.
78. N. Savvides and B. Window, "Unbalanced Magnetron Ion-Assisted Deposition and Property Modification of Thin Films," *J. Vac. Sci. Technol.*, **A4**(3), 504, 1986.
79. W. D. Sproul, P. J. Rudnik, M. E. Graham, and S. L. Rohde, "High rate reactive sputtering in an opposed cathode closed-field unbalanced magnetron sputtering system," *Surf. Coat. Technol.*, **43/44**, 270, 1990.

80. S. L. Rohde, I. Petrov, W. D. Sproul, S. A. Barnett, P. J. Rudnik, and M. E. Graham, "Effects of an Unbalanced Magnetron in a Unique Dual-Cathode, High Rate Reactive Sputtering System," *Thin Solid Films*, **193/194**, 117, 1990.
81. K. Tominaga, "Preparation of AlN Films by Planar Magnetron Sputtering with Facing Two Targets," *Vacuum*, **41**, 1154, 1990.
82. S. Kadlec, J. Musil, V. Valvoda, W.-D. Munz, H. Pewtersei, and J. Schroeder, "TiN Films Grown by Reactive Magnetron Sputtering with Enhanced Ionisation at Low Discharge Pressures," *Vacuum*, **41**(7-9), 2233, 1990.
83. S. Kadlec, J. Musil, and W.-D. Munz, "Sputtering Systems with Magnetically Enhanced Ionization for Ion Plating of TiN Films," *J. Vac. Sci. Technol.*, **A8**(3), 1318, 1990.
84. R. P. Howson, H. A. J'Afer, and A. Spencer, "Substrate Effects from an Unbalanced Magnetron," *Thin Solid Films*, **193/194**, 127, 1990.
85. R. P. Howson and H. A. J'Afer, "Reactive Sputtering with an Unbalanced Magnetron," *J. Vac. Sci. Technol.*, **A10**(4), 1784, 1992.
86. C. F. Morrison, U.S. Patent No. 4,351,472, 1982.
87. C. F. Morrison and R. P. Welty, *Anodic Plasma Generation in Magnetron Sputtering*, Vac-Tec Systems, Boulder, CO, 1982.
88. D. Hofman, S. Beisswanger, and A. Feuerstein, "Novel Low Temperature Hard Coatings for Large Parts," *Surf. Coat. Technol.*, **49**, 330, 1991.
89. T. Leyendecker, O. Lemmer, S. Esser, and J. Ebberink, "The Development of the PVD Coating TiAlN as a Commercial Coating for Cutting Tools," *Surf. Coat. Technol.*, **48**, 175, 1991.
90. S. M. Rossnagel and J. Hopwood, "Metal Ion Deposition from Ionized Magnetron Sputtering Discharge," *J. Vac. Sci. Technol.*, **B12**(2), 449, 1994.
91. J. M. Schneider, W. D. Sproul, and A. Matthews, "Reactive Ionized Magnetron Sputtering of Crystalline Alumina Coatings," *Surf. Coat. Technol.*, **98**, 1473, 1998.
92. J. M. Schneider, W. D. Sproul, A. A. Voevodin, and A. Matthews, "Crystalline Alumina Deposited at Low Temperatures by Ionized Magnetron Sputtering," *J. Vac. Sci. Technol.*, **A15**(3), 1084, 1997.
93. A. Blondeel, P. Persoone, and W. De Bosscher, *Vakuum in Forschung und Praxis (VIP)*, **21**(3), 6, 2009.
94. J. Lehan, H. Byorum, R. J. Hill, and J. Kirkwood Rough, U.S. Patent No. 5,814,195, 1998.
95. K. F. Lai, K. Song, and D. B. Hayden, U.S. Patent No. 7585,399, 2009.
96. G. Dearnaley, in D. S. Rickerby and A. Matthews (Eds.), *Advanced Surface Coatings: A Handbook of Surface Engineering*, Blackie, London, 1991.
97. A. Matthews, "Plasma Assisted PVD: The Past and Present," *SVC Bull.* **24**, Fall 2013.
98. G. K. Hubler and J. K. Hirvonen, *ASM Handbook*, Vol. 5: *Surface Engineering*, ASM International, Metals Park, OH, 1994.
99. P. J. Martin, R. P. Netterfield, and W. G. Sainty, "Modification of the Optical and Structural Properties of Dielectric ZrO₂ Films by Ion-Assisted Deposition," *J. Appl. Phys.*, **55**, 235, 1984.
100. P. J. Martin, H. A. Macleod, R. P. Netterfield, C. G. Pacey, and W. G. Sainty, "Ion-Beam-Assisted Deposition of Thin Films," *Appl. Opt.*, **22**, 178, 1983.
101. F. L. Williams, R. D. Jacobson, J. R. McNeil, G. J. Exarhos, and J. J. McNally, "Optical Characteristics of Thin Films Deposited at Low Temperature Using Ion Assisted Deposition," *J. Vac. Sci. Technol.*, **A6**, 2020, 1988.
102. G. K. Hubler, C. A. Carosella, P. G. Burkhalter, R. K. Feitag, C. M. Cotell, and W. D. Coleman, "Fabrication of Low-Z X-Ray Mirrors by Ion Beam Assisted Deposition," *Nucl. Instrum. Methods Phys. Res. B*, **59/60**, 268, 1991.
103. J. D. Targove, L. J. Lingg, J. P. Lecham, C. K. Hwangbo, H. A. Macleod, J. A. Leavitt, and L. C. McIntyre, Jr., in U. J. Gibson, A. E. White, P. P. Pronko (Eds.), *Materials Modification and Growth Using Ion Beams*; Symposium held 21-23 April 1987, Anaheim, CA, Materials Research Society, Pittsburgh, PA, 1987.
104. R. A. Kant, S. A. Dillich, B. D. Sartwell, and J. A. Sprague, "The Causes of Property Variations of ibad-Titanium Nitride," *Proc. Mater. Res. Soc. Symp.*, **128**, 427, 1989.

105. W. G. Sainty, P. J. Martin, R. P. Netterfield, D. R. McKenzie, D. J. H. Cockayne, and D. M. Dwarde, "The Structure and Properties of Ion-Beam-Synthesized Boron Nitride Films," *J. Appl. Phys.*, **64**, 3980, 1988.
106. S. Pimbert-Michaux, C. Chabrol, M. F. Denanot, and J. Delafond, "Structure and Properties of Titanium Carbide Grown by Dynamic Ion Beam Mixing," *Mater. Sci. Eng. A*, **115**, 209, 1989.
107. A. Dehbi-Alaoui, A. Matthews, and J. Franks, "The Optical and Mechanical Properties of Carbon Films Grown Using a Fast Atom Beam Source," *Surf. Coat. Technol.*, **47**, 722, 1991.
108. A. A. Voevodin, J. M. Schneider, P. Stevenson, and A. Matthews, "Studies of Atom Beams Produced by a Saddle Field Source Used for Depositing Diamond-Like Carbon Films on Glass," *Vacuum*, **46**, 299, 1995.
109. J. C. Angus, P. Koidl, and S. Domitz, in J. Mart and F. Jansen (Eds.), *Plasma Deposited Thin Films*, CRC, Boca Raton, FL, 1986.
110. P. J. Martin, R. P. Netterfield, A. Bendavid, and T. J. Kinder, "The Deposition of Thin Films by Filtered Arc Evaporation," *Surf. Coat. Technol.*, **54/55**, 136, 1992.
111. V. N. Zhitomirsky, R. L. Boxman, and S. Goldsmith, "Influence of an External Magnetic Field on Cathode Spot Motion and Coating Deposition Using Filtered Vacuum Arc Evaporation," *Surf. Coat. Technol.*, **68/69**, 146, 1994.
112. D. R. McKenzie, D. Muller, and B. A. Pailthorpe, "Compressive-Stress-Induced Formation of Thin-Film Tetrahedral Amorphous Carbon," *Phys. Rev. Lett.*, **67**, 773, 1991.
113. A. A. Voevodin and M. S. Donley, "Preparation of Amorphous Diamond-Like Carbon by Pulsed Laser Deposition: A Critical Review," *Surf. Coat. Technol.*, **82**, 199, 1996.
114. S. V. Prasad, J. S. Zabinski, and N. T. McDevitt, "Friction Behaviour of Pulsed Laser Deposited Tungsten Disulphide Films," *STLE Tribol. Trans.*, **38**, 57, 1995.
115. A. A. Voevodin, M. A. Capano, A. J. Safriet, M. S. Donley, and J. S. Zabinski, "Combined Magnetron Sputtering and Pulsed Laser Deposition of Carbides and Diamond-Like Carbon Films," *Appl. Phys. Lett.*, **69**, 188, 1996.
116. A. Leyland and A. Matthews, "On the Significance of the H/E Ratio in Wear Control: A Nanocomposite Coating Approach to Optimal Tribological Behaviour," *Wear*, **246**, 1, 2000.
117. A. Leyland and A. Matthews, "Optimisation of Nanostructured Tribological Coating," in J. T. M. de Hosson and A. Cavaleiro (Eds), *Hard Nanostructured Coatings*, Springer, 2006, Chapter 12.
118. A. Matthews and A. Leyland, "Materials Related Aspects of Nanostructured Tribological Coatings," *SVC Bull.*, Spring, 40, 2009.
119. G. Cassar, A. Matthews, and A. Leyland, "Triode Plasma Diffusion Treatment of Titanium Alloys," *Surf. Coat. Technol.*, **212**, 20, 2012.

CHAPTER 8

MECHANICAL FASTENERS

Murray J. Roblin
California State Polytechnic University
Pomona, California

Updated by Anthony Luscher
The Ohio State University
Columbus, Ohio

1 INTRODUCTION TO FASTENING AND JOINING	255	8.2 Torsional Stress Factor	271
1.1 Assembly Features and Functions	256	8.3 Other Design Issues	271
1.2 Some Examples of a Nesting Strategy	259	9 THEORETICAL BEHAVIOR OF THE JOINT UNDER TENSILE LOADS	272
1.3 Three-Part Assembly	260	9.1 Critical External Load Required to Overcome Preload	274
2 INTRODUCTION TO FASTENING WITH BOLTS AND RIVETS	261	9.2 Very Large External Loads	275
3 BOLTED AND RIVETED JOINT TYPES	262	10 EVALUATION OF SLIP CHARACTERISTICS	276
4 EFFICIENCY	264	11 TURN-OF-NUT METHOD OF BOLT TIGHTENING	277
5 STRENGTH OF A SIMPLE LAP JOINT	264	12 TORQUE AND TURN TOGETHER	278
6 SAMPLE PROBLEM OF A COMPLEX BUTT JOINT (BEARING-TYPE CONNECTION)	266	13 ULTRASONIC MEASUREMENT OF BOLT STRETCH OR TENSION	279
6.1 Preliminary Calculations	266	14 FATIGUE FAILURE AND DESIGN FOR CYCLICAL TENSION LOADS	280
7 FRICTION-TYPE CONNECTIONS	268	REFERENCES	282
8 UPPER LIMITS ON CLAMPING FORCE	271		
8.1 Design-Allowable Bolt Stress and Assembly Stress Limits	271		

1 INTRODUCTION TO FASTENING AND JOINING

The study of fastening and joining is complex but worthy of study because of the economic, structural, reliability, safety, and structural efficiency benefits that can be achieved. It is also an ever-changing field of study with advances constantly being made. As this document goes to

press in 2014, Boeing is developing a commercial aircraft with will be a step increment in lower cost, comfort, and fuel efficiency. These goals could only be achieved by use of a full composite structure making extensive use of adhesives along precisely controlled bond lines. As a general reference for the study of assembly, I recommend *Mechanical Assembly* by Daniel Whitney.¹

The technical area of fastening and joining is a very wide area drawing from many engineering disciplines. It comprises a great deal of specialized knowledge within each joining and fastening methodology. A complete survey of all of these areas would require a series of texts and is beyond the scope of this chapter. Instead this chapter will focus on mechanical fastening via rivets and bolts. Throughout this chapter we use the term *fastening* to correspond to removable features and *joining* as the creation of a permanent connect.

Before these mechanical fastening systems are discussed in detail, a section on common assembly issues is presented. This section discusses how a creative consideration of assembly constraint can improve the design of mechanically fastened assemblies.

1.1 Assembly Features and Functions

To successfully join parts and form an assembly four primary tasks must be accomplished:

- Location of the parts relative to each other. Surfaces must contact to remove all degrees of freedom between the parts except those to be removed by the final locking features. Location is not complete until the part is fully constrained with respect to moments and translations.
- Transfer of service loads across the interfaces of the assembly. These are commonly the same features used to form location but need not be. They must, however, have sufficient strength and rigidity to transfer load.
- If necessary, part tolerance and manufacturing variability between the parts must be absorbed by use of shims or compliance features. This is not necessary in all assemblies.
- Addition of locking features such as bolts to finish the constraint of the parts to each other.

Although fastening can be done creatively in 3-space, this is not often done. The current situation is called by the author the *bolting paradigm* and can be characterized by

- Using assembly interfaces that are two dimensional and predominately planar.
- Having all bolts take loads in multiple directions (axial, shear, moment).
- Tolerancing is not a concern as long as the bolt can be assembled through the mating parts.

Figure 1 shows two examples. Another approach to joining parts, which can provide great benefit in loading and cost, is to use a more complex three-dimensional assembly interface. In this strategy,

- Individual features take loads only in specific direction.
- Assemblies can be designed to be statically determinate or with various levels of over-constraint.
- Assembly features and not the fasteners determine tolerances.

To create such a three-dimensional fastening strategy, surfaces or features must be created to accomplish three assembly functions: locators, compliant features, and locks.

Locators are features or mating surfaces that eliminate degrees of freedom between parts, transfer the service loads, and/or establish the major reference or datum planes or points that

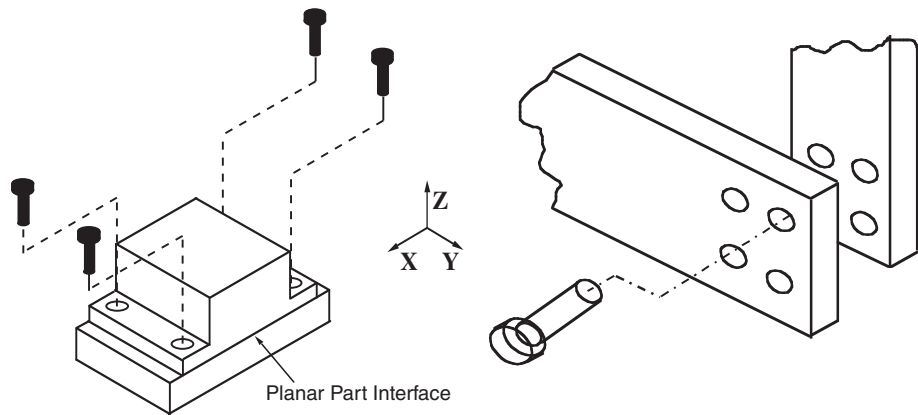


Figure 1 Examples of the two-dimensional bolting paradigm.

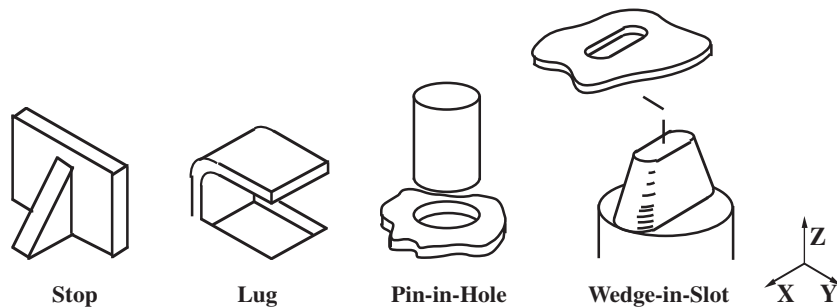


Figure 2 Several features that provide locational constraint between parts.

locate parts relative to each other. For two parts to remain together as an assembly and function, their relative location, alignment, and orientation must be fixed for all time.

Locators are available in many different geometries and topologies. They can be designed for very specific loading situations or for general situations. When designing with locators, it is useful to consider the degrees of freedom eliminated by each feature or surface pair. Several features that can be machined, molded, or bent into a part are shown in Fig. 2.

To create this type of assembly, start the process by considering one part of the assembly as the reference frame and fixed to ground. Then assume that the other part (the mating part) starts out with all degrees of freedom available to it. Normally a part would have 6 possible degrees of motion in 3-space (for example, 3 Cartesian coordinate translations and 3 rotations). Physical assembly is somewhat more difficult than this since it is possible to constrain an object in only one direction along an axis. The simplest example of this is a planar surface that constrains motion into the surface but not away from it. Therefore, both positive and negative directions need to be considered separately, leading to 12 possible degrees of motion that are available between two parts (6 Cartesian coordinate translations and 6 rotations).

For example, the “stop” feature constrains only motion that is inward normal to its large surface (in the positive x direction in Fig. 2). It does not constrain motion in any other direction or outwardly normal from its large surface. The “pin-in-hole” feature, on the other hand, constrains all planar motion normal to its centerline in both x and y directions. A pin-in-hole feature,

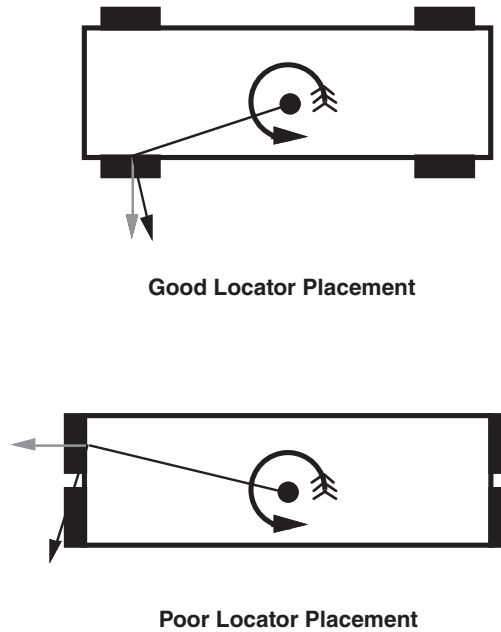


Figure 3 Importance of line of action in the use of locators.

however, will allow in-plane rotation around the pin. The “wedge-in-slot” feature limits motion to being along the slot. In forming these locator definitions it is important to include surfaces on both parts since they both contribute to location functionality. As an example, defining a pin feature is not sufficient to determine the degrees of freedom that are eliminated. A pin-in-hole feature will remove different degrees of freedom than a pin-in-slot feature.

Note that several different physical features can be used to achieve identical locator functionality. Because these features are physically different, they will differ in moldability, strength, and ability to absorb manufacturing variability. These secondary attributes should be considered after determining the degrees of freedom that need to be removed. In this way an assembly concept can progress from a very abstract level to a more physical level in a logical manner.

Figure 3 shows another important aspect of using locators, which is their line of action in providing constraint. In Fig. 3 assume that the rectangular box is being constrained against rotation by the four stop features shown surrounding it. In these graphics the gray arrows represent the normal to the constraint feature while the black arrow shows the force couple caused by the imposed moment. Note that in the top graphic the gray and black arrows are almost colinear, while in the bottom graphic they are almost orthogonal.

Compliant features are designed to absorb any tolerance stack-up or misalignment between the data of mating parts. This is most often accomplished by the built-in compliance or flexibility of these features.

All parts are manufactured with some form of variability or tolerance. This variability occurs in many forms, including errors of size, location, orientation, and form. Tolerance stack-up can (a) produce a gap between the parts resulting in an undesirable rattle or looseness or (b) produce an unintended interference leading to high stresses in the part and high assembly forces. The feature group “compliant features” eliminates any resultant gap or interferences

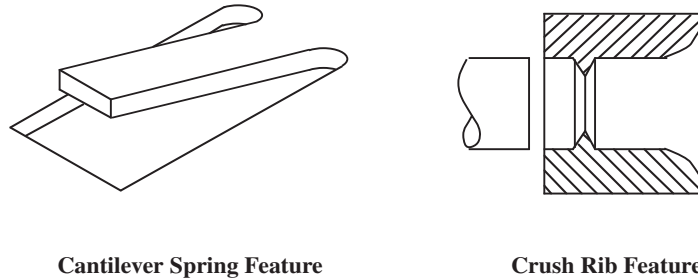


Figure 4 Example of an elastic and an inelastic compliance feature.

between the data of mating parts by building in a certain amount of flexibility or compliance between the parts. Compliant features function in one of the following two ways:

- *Elastically.* These features are designed so that, under all tolerance stack-ups, a guaranteed minimum amount of preload is maintained in the system. The left graphic in Fig. 4 shows a common feature used in the sheetmetal industry. The spring finger's undeformed height is above the surrounding surface. When parts are assembled, the finger elastically deforms but maintains pressure between the parts.
- *Inelastically.* These features are designed so that, during initial assembly, they permanently deform to eliminate any gaps between parts. The right graphic in Fig. 4 shows a crush rib that is often used in plastic parts. In this design a steel shaft needs to have a tight fit against a polymer housing. The first time that the shaft is inserted the rib is fractured to the exact diameter needed. Another example from metals is a crush washer, which is used to deform material and provide a sealing surface.

The compliant feature classification is important since it represents an inexpensive method of compensating for variability versus the very high cost of tightening tolerances in both the base and mating parts.

Locks are features or devices used to provide the final attachment between two parts. Locks can be any of a wide spectrum of fasteners, including bolts, rivets, or snap-fits. It is important to note that whatever lock feature or fastener is chosen, its kinematic role in the overall structure is the same.

1.2 Some Examples of a Nesting Strategy

The use of the above feature types can best be shown by an example. Figure 5 is a schematic of two ways of attaching a transformer base to a sheet-metal panel. The graphic on the left shows a conventional way of fastening the structure by use of four fasteners. This approach, however, has the disadvantage of requiring four fasteners, and also can lead to tolerance issues between the four screw holes. The concept on the right shows an approach that uses the selective constraint of locating features. The sheet-metal is bent to form a three-sided support structure that is kinematically equivalent to the lug feature shown in Fig. 2. They constrain the vertical direction as well as motion toward the back surface. By the use of three sets of these features the only allowable motion is shown in Fig. 5. More importantly, only one locking feature is required, and it needs to provide constraint only in the opposite direction of the assembly motion. Tolerance windows can be opened and are set by the dimensions of the locators. If rattle needs to be eliminated, compliance features could have been added.

Figure 6 shows another example. It is the attachment of a lens to a bulb housing. This example uses three assembly feature types and is from an automotive interior. Most of the

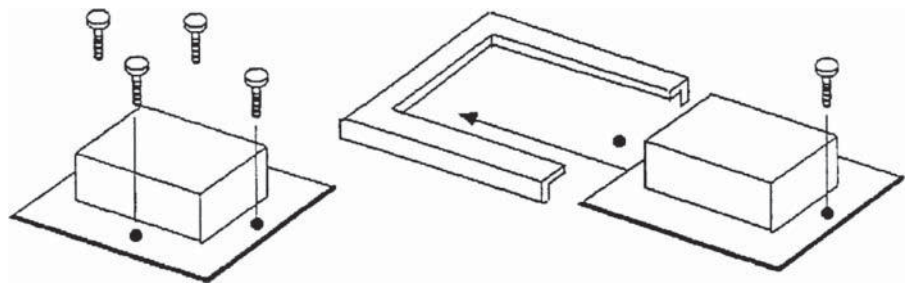


Figure 5 Transformer attached to a sheetmetal panel.

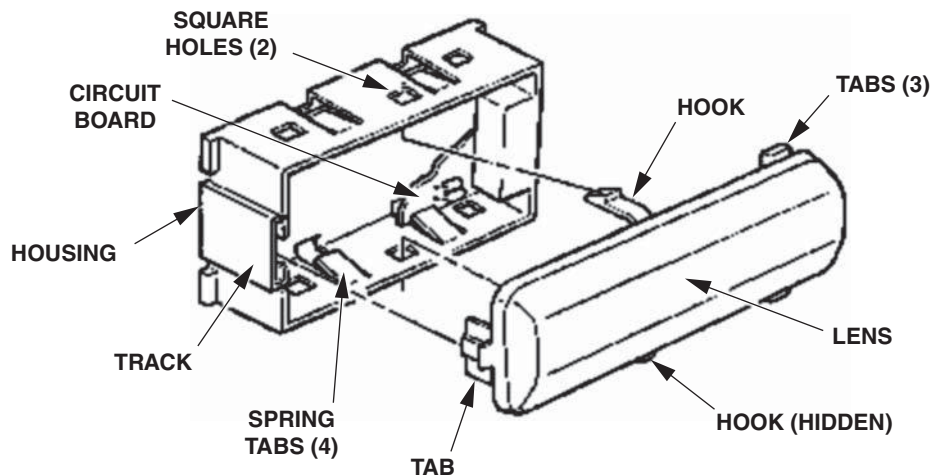


Figure 6 Lens to bulb housing assembly example.

constraint between the parts is provided by the mating of the track and tab features on the housing and lens, respectively. These two features remove all degrees of freedom except motion along and away from the assembly direction. The back of the lens hitting the front of the housing removes motion along the assembly direction. Motion away from the assembly motion is removed by the locking device, which in this case, is a snap-fit. Since rattle is a concern in an automotive environment, four spring tabs are the most cost-effective solution to providing some preload between the parts.

1.3 Three-Part Assembly

In many cases fastening efficiencies can be improved by the creation of what the author calls three-part assemblies. In this assembly strategy a part can be sandwiched between two other parts that are fastened together. In the left-hand graphic in Fig. 7, the handle is attached to the top part while fasteners on the top part connect it to the bottom part. Consider the graphic on the right side below. A single set of fasteners is used to connect the handle, top, and bottom parts together. The top part is completely constrained and no additional fasteners are needed. Other examples of multipart assemblies include gears in a gear housing. Often several gears are in parallel bores. They are all constrained together when the two housings are fastened together.

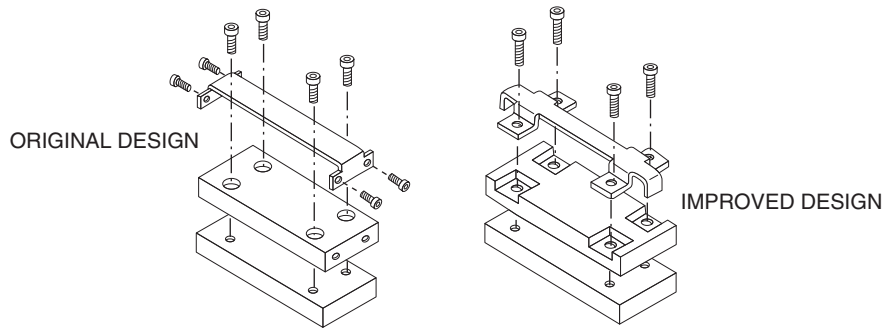


Figure 7 Three-part assembly.

2 INTRODUCTION TO FASTENING WITH BOLTS AND RIVETS

Two or more components may need to be fastened in such a way that they may be taken apart during the service life of the part. In these cases, the assembly must be fastened mechanically. Other reasons for choosing mechanical fastening over welding could be

1. Ease of part replacement, repair, or maintenance
2. Ease of manufacture
3. Designs requiring movable joints
4. Designs requiring adjustable joints

The most common mechanical fastening methods are bolts (threaded fasteners) and rivets.

To join two members by bolting or riveting requires holes to be drilled in the parts to accommodate the rivets or bolts. These holes reduce the load-carrying cross-sectional area of the members to be joined. Because this reduction in area is at least 10–15%, the load-carrying capacity of the bolted structure is reduced, which must be accounted for in the design. Alternatively, when one inserts bolts into the holes, only the cross section of the bolt or rivet supports the load. In this case, the reduction in the strength of the joint is reduced even further than 15%.

Even more critical are the method and care taken in drilling the holes. When one drills a hole in metal, not only is the cross-sectional area reduced, but the hole itself introduces stress risers and/or flaws on the surface of the holes that may substantially endanger the structure. First, the hole places the newly created surface in tension, and if any defects are created as a result of drilling, they must be accounted for in a quantitative way. Unfortunately, it is very difficult to obtain definitive information on the inside of a hole that would allow characterization of the introduced defect.

The only current solution is to make certain that the hole is properly prepared, which means not only drilling or subpunching to the proper size but also *reaming* the surface of the hole. To be absolutely certain that the hole is not a problem, one needs to put the surface of the hole in residual compression by expanding it slightly with an expansion tool or by pressing the bolt, which is just slightly larger than the hole. This method causes the hole to expand during insertion, creating a hole whose surface is in residual compression. While there are fasteners designed to do this, it is not clear that all of the small surface cracks of the hole have been removed to prevent flaws or stress risers from existing in the finished product.

Using bolts and rivets in an assembly can also provide an ideal location for water to enter the crevices between the two joined parts. This trapped water, under conditions where chlorides and sodium exist, can cause *crevice corrosion*, which is a serious problem if encountered.

Obviously, in making the holes as perfect as possible, you increase the cost of a bolted and/or riveted joint significantly, which makes welding or adhesive joining a more attractive option. Of course, as will be shown below, welding and joining have their own set of problems that can degrade the joint strength.

The analysis of the strength of a bolted or riveted joint involves many indeterminate factors resulting in inexact solutions. However, by making certain simplifying assumptions, we can obtain solutions that are conservative, acceptable, and practical. We discuss two types of solutions: *bearing-type connections*, which use ordinary or unfinished bolts or rivets, and *friction-type connections*, which use high-strength bolts. Today, economy and efficiency are obtained by using high-strength bolts for field connections together with welding in the shop. With the advent of lighter-weight welding power supplies, the use of field welding combined with shop welding is finding increasing favor.

While riveted joints do show residual clamping forces (even in cold-driven rivets), the clamping forces in the rivet is difficult to control, is not as great as that developed by high-strength bolts, and cannot be relied upon. Hot driven rivets have fallen out of favor due to the cost and safety issues involved. Most commercial and bridge structures are now designed using bolts. Studies have shown that the holes are almost completely filled for short rivets. As the grip length is increased, the clearances between rivet and plate material tend to increase.

3 BOLTED AND RIVETED JOINT TYPES

There are two types of riveted and bolted joints: *lap joints* and *butt joints*. See Figs. 8 and 9 for lap and butt joints, respectively. Note that there can be one or more rows of connectors, as shown in Fig. 9a and b.

In a butt joint, plates are butted together and joined by two cover plates connected to each of the main plates. (Rarely, only one cover plate is used to reduce the cost of the joint.) The number of rows of connectors that fasten the cover plate to each main plate identifies the joint—single row, double row, and so on. See Fig. 9.

Frequently the outer cover plate is narrower than the inner cover plate, as in Fig. 9c and d, the outer plate being wide enough to include only the row in which the connectors are most closely spaced. This is called a *pressure joint* because caulking along the edge of the outer cover plate to prevent leakage is more effective for this type of joint.

The spacing between the connectors in a given row is called the *pitch*. When the spacing varies in different rows, as in Fig. 9d, the smallest spacing is called the *short pitch*, the next smallest the *intermediate pitch*, and the greatest the *long pitch*. The spacing between consecutive rows of connectors is called the *back pitch*. When the connectors (rivets or bolts) in consecutive rows are staggered, the distance between their centers is the *diagonal pitch*.

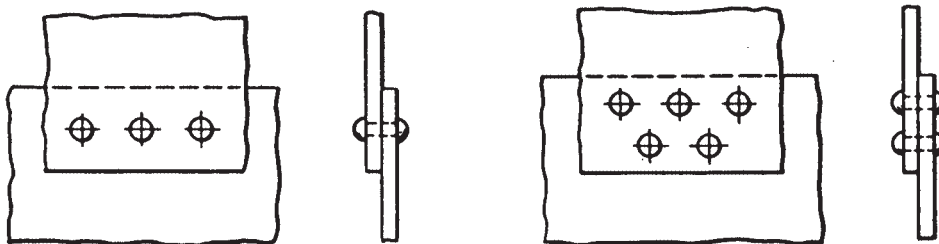


Figure 8 Lap joints. Connectors are shown as rivets only for convenience.

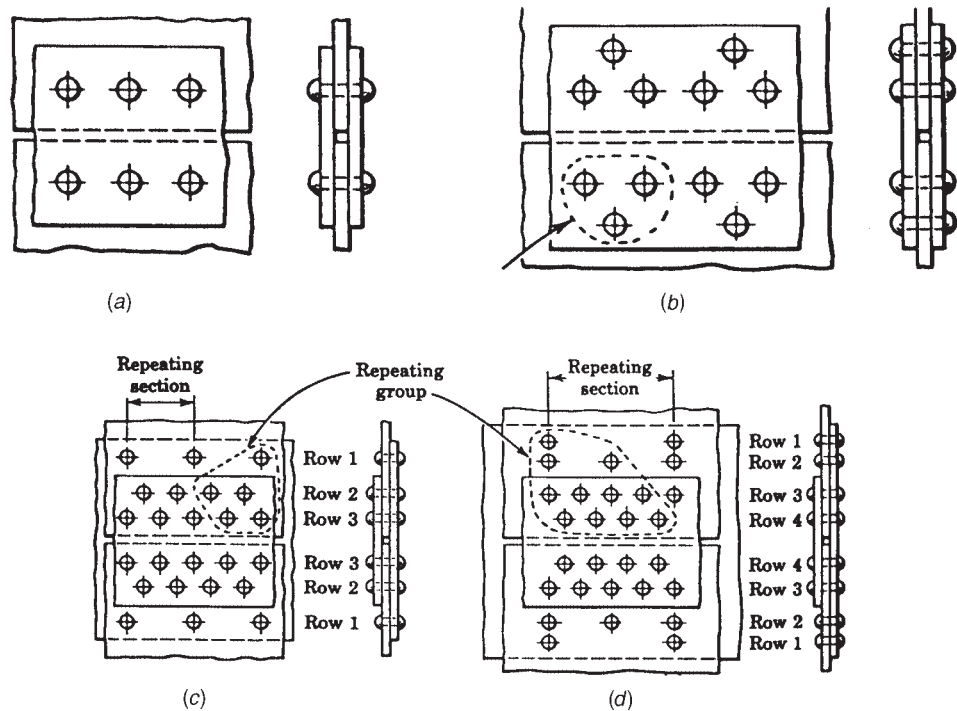


Figure 9 Butt joints: (a) single row, (b) double row, (c) triple row (pressure type); and (d) quadruple row (pressure type).

In determining the strength of a joint, computations are usually made for the length of a joint corresponding to a repeating pattern of connectors. The length of the repeating pattern, called the *repeating section*, is equal to the long pitch.

To clarify how many connectors belong in a repeating section, see Fig. 9c, which shows that there are 5 connectors effective in each half of the triple row—that is, 2 half connectors in row 1, 2 whole connectors in row 2, and 1 whole and 2 half connectors in row 3. Similarly, there are 11 connectors effective in each half of the repeating section in Fig. 9d.

When rivets are used in joints, the holes are usually drilled or punched and reamed out to a diameter of $1/16$ in. (1.5 mm) larger than the nominal rivet size. The rivet is assumed to be driven so tightly that it fills the hole completely. Therefore, in calculations the diameter of the hole is used because the rivet fills the hole. This is not true for a bolt unless it is very highly torqued. In this case, a different approach needs to be taken, as delineated later in this chapter.

The fastener patterns shown in Fig. 9b–d are designed with multiple rows in order to spread the load out among the different fasteners. The number of fasteners in each row is different and is determined by the concept of elastic matching. In this concept, the goal is to carry the same load in each fastener to the greatest extent possible.

Consider the three-row riveted joint as shown in Fig. 9c. Notice that this is a statically indeterminate structure and the only way to solve for forces is to consider the stiffness of each load path. Each row of rivets forms a different load path and has a different stiffness due to the different distances from the center of the joint. Row 3, for example, has the shortest distance between its counterpart on each side of the joint. Because of this, the load path for row 3 is the stiffest of the three load paths. Row 1, on the other hand, has the longest load path, the greatest length of plate that can deform, and therefore the lowest stiffness of the three.

When solving statically indeterminate structures, load is distributed as the ratio of relative stiffness. As an example, if a load has three paths, the path with the highest stiffness will take the highest load. In the three-rowed joint, row 3 takes the greatest load since it is the stiffest. Row 2 has an intermediate stiffness, and row 1 has the lowest stiffness. In order to equalize the load on each fastener, the row with the highest load (row 3) should have the greatest number of fasteners, and row 1 should have the least. With an exact calculation of stiffness the loads can be fine-tuned for joint efficiency.

4 EFFICIENCY

Efficiency compares the strength of a joint to that of a continuous solid plate as follows:

$$\text{Efficiency} = \frac{\text{Strength of the joint}}{\text{Strength of solid plate}}$$

5 STRENGTH OF A SIMPLE LAP JOINT

For bearing-type connections using rivets or ordinary bolts, we first consider failure in the bolt or rivet itself in shear. This leads to the following equation:

$$P_s = A_s \tau$$

Using the diameter of the bolt/rivet, this can be rewritten as

$$P_s = A_s t = \frac{pd^2 t}{4}$$

where P_s = load

A_s = shear area of one connector

d = diameter of connector and/or hole

For the above example, friction is neglected. Figure 10 shows the shearing of a single connector.

Another possible type of failure is caused by tearing the main plate. Figure 11 demonstrates this phenomenon.

The above failure occurs on a section through the connector hole because this region has reduced tearing resistance. If p is the width of the plate or the length of a repeating section, the resisting area is the product of the net width of the plate ($p - d$) times the thickness t . The failure load in tension therefore is

$$P_{\text{tension}} = A_t \sigma_t = (p - d)t(\sigma_t)$$

A third type of failure, called a *bearing failure*, is shown in Fig. 12 For this case, the edge of the plate yields and the bolt hole is enlarged into a slot. Actually, the stress that the

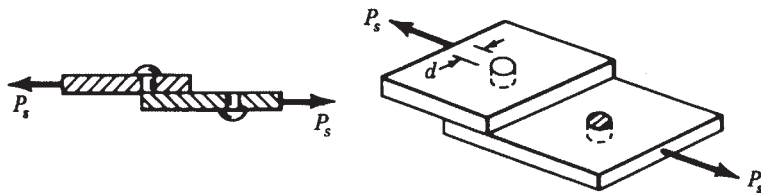


Figure 10 Shear failure.

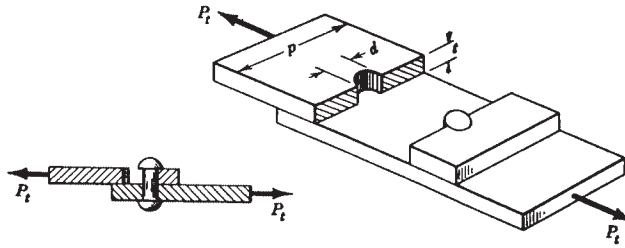


Figure 11 Tear of plate at section through connector hole. $P_t = A_t \sigma_t = (p - d)t\sigma_t$.

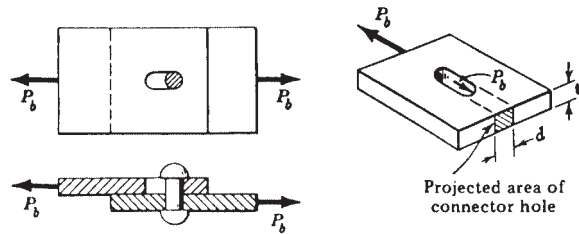


Figure 12 Exaggerated bearing deformation of upper plate. $P_b = A_b \sigma_b = (td)\sigma_b$.

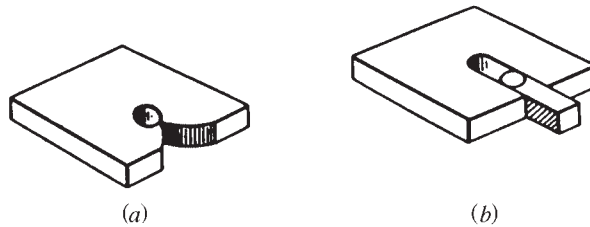


Figure 13 Possible types of failure if connector hole is too close to edge of plate: (a) tear out and (b) shear behind connector.

connector bears against the edges of the hole varies from zero at the edges of the hole to the maximum value at the center of the bolt or rivet. However, common practice assumes the stress as uniformly distributed over the projected area of the hole. See Fig. 12.

The failure load in the bearing area can be expressed by

$$P_b = A_b \sigma_b = (td)\sigma_b$$

Other types of failure are possible but will not occur in a properly designed joint. These are tearing of the edge of the plate back of the connector hole (Fig. 13a) or a shear failure behind the connector hole (Fig. 13b) or a combination of both. Failures of this type will not occur when the distance from the edge of the plate is greater than two times the diameter of the bolt/rivet.

6 SAMPLE PROBLEM OF A COMPLEX BUTT JOINT (BEARING-TYPE CONNECTION)

The strength of a bearing-type connection is limited by the capacity of the rivets or ordinary bolts to transmit load between the plates or by the tearing resistance of the plates themselves, depending on which is smaller. The calculations are divided as follows:

1. Preliminary calculations to determine the load that can be transmitted by one rivet or bolt in shear or bearing *neglecting friction* between the plates
2. Calculations to determine which mode of failure is most likely

A repeating section 180 mm long of a riveted triple row butt joint of the pressure type is illustrated in Fig. 14. The rivet hole diameter $d = 20.5$ mm, the thickness of the main plate $t = 14$ mm, and the thickness of each cover plate $t' = 10$ mm. The ultimate stresses in shear, bearing, and tension are, respectively, $\tau = 300$ MPa, $\sigma_b = 650$ MPa, and $\sigma_t = 400$ MPa. Using a factor of safety of 5, determine the strength of a repeating section, the efficiency of the joint, and the maximum internal pressure that can be carried in a 1.5-m-diameter boiler where this joint is the longitudinal seam.

Solution. The use of ultimate stresses will determine the ultimate load, which is then divided by the factor of safety (in this case 5) to determine the safe working load. An alternative but preferable procedure is to use allowable stresses to determine the safe working load directly, which involves smaller numbers. Thus, dividing the ultimate stressed by 5, we find that the allowable stresses in shear, bearing, and tension, respectively, are $\tau = 300/5 = 60$ MPa, $\sigma_b = 650/5 = 130$ MPa, and $\sigma_t = 400/5 = 80$ MPa. The ratio of the shear strength τ to the tensile strength σ of a rivet is about 0.75.

6.1 Preliminary Calculations

To single shear one rivet,

$$P_s = \frac{pd^2}{4}\tau = \frac{p}{4}(20.5 \times 10^{-3})^2(60 \times 10^6) = 19.8 \text{ kN}$$

As shown in the bottom of Fig. 14, to move the main plates the rivets must be sheared in two places. To double shear one rivet,

$$P_s = 2 \times 19.8 = 39.6 \text{ kN}$$

To have bearing failure in one rivet in the main plate,

$$P_b = (td)\sigma_b = (14.0 \times 10^{-3})(20.5 \times 10^{-3})(130 \times 10^6) = 37.3 \text{ kN}$$

To crush one rivet in one cover plate,

$$P'_b = (t'd)\sigma_b = (10 \times 10^{-3})(20.5 \times 10^{-3})(130 \times 10^6) = 26.7 \text{ kN}$$

Rivet Capacity Solution. The strength of a single rivet in row 1 in a repeating section is determined by the lowest value of the load that will single shear the rivet, crush it in the main plate, or crush it in one of the cover plates. Based on the values in the preceding calculations, this value is 19.8 kN per rivet.

The strength of each of the two rivets in row 2 depends on the lowest value required to double shear the rivet, crush it the main plate, or crush it in both cover plates. From the above preliminary calculations, this value is 37.3 kN per rivet or $2 \times 37.3 + 74.6$ kN for both rivets in row 2.

Each of the two rivets in the repeating section in row 3 transmits the load between the main plate and the cover plate in the same manner as those in row 2; hence for row 3, the strength = 74.6 kN.

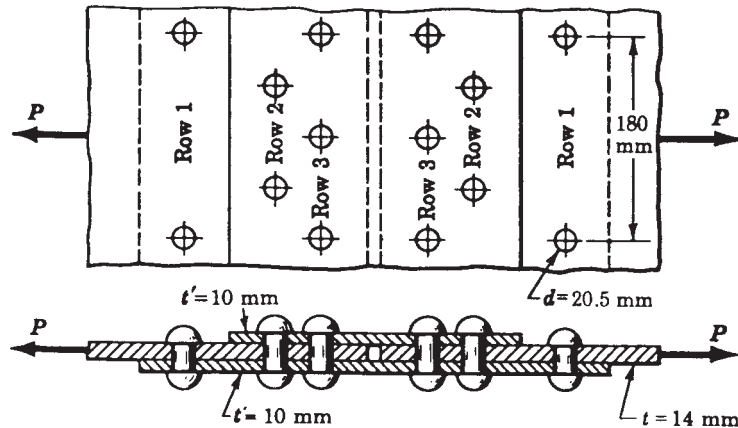


Figure 14

The total rivet capacity is the sum of the rivet strengths in all rows (rows 1, 2, 3), as follows:

$$P_{\text{total}} = 19.8 + 74.6 + 74.6 = 169.0 \text{ kN}$$

Tearing Capacity. The external load applied to the joint acts directly to tear the main plate at row 1, and the failure would be similar to Fig. 11. This is calculated as follows:

$$P_{\text{tearing}} = (p - d)\sigma_t = [(180 \times 10^{-3}) - (20.5 \times 10^{-3})](14 \times 10^{-3})(80 \times 10^6) = 178.6 \text{ kN}$$

The external load applied does not act directly to tear the main plate at row 2 because part of the load is absorbed or transmitted by the rivet in row 1. Hence, if the main plate is to tear at row 2, the external load must be the sum of the tearing resistance of the main plate at row 2 plus the load transmitted by the rivet in row 1. See Figs. 15 and 16.

Thus,

$$\begin{aligned} P_{\text{tearing2}} &= (p - 2d)t\sigma_t + \text{Rivet strength in row 1} \\ &= [(180 \times 10^{-3}) - 2(20.5 \times 10^{-3})](14 \times 10^{-3})(80 \times 10^6) + 19.8 \times 10^3 \\ &= 175.5 \text{ kN} \end{aligned}$$

Similarly, the external load required to tear the main plate at row 3 must include the rivet resistance in rows 1 and 2 or

$$\begin{aligned} P_3 &= [(180 \times 10^{-3}) - 2(20.5 \times 10^{-3})](14 \times 10^{-3})(80 \times 10^6) + (19.8 \times 10^3) + (74.6 \times 10^3) \\ &= 250.1 \text{ kN} \end{aligned}$$

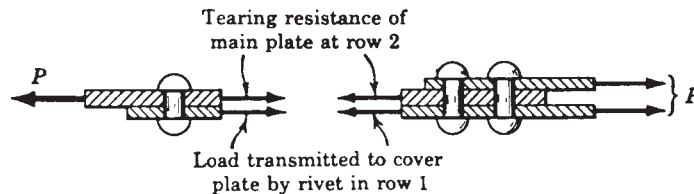


Figure 15

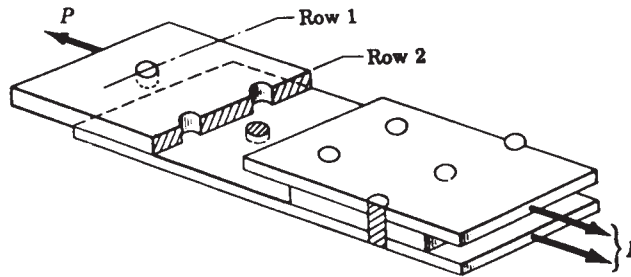


Figure 16 Failure by shear of rivet in row 1 plus tear of main plate in row 2.

It is obvious that this computation need not be made because the tearing resistance of the main plates at rows 2 and 3 is equal, thus giving a larger value.

At row 3, the tearing resistance of the cover plates is resisted by the tensile strength of the reduced section of that row. The tensile strength of one cover plate is

$$P_c = [(180 \times 10^{-3}) - 2(20.5 \times 10^{-3})](10 \times 10^{-3})(80 \times 10^6) = 111.2 \text{ kN}$$

In an ordinary butt joint, the tensile capacity of both cover plates is twice this value. In a pressure joint, however, where one cover plate is shorter than the other, the load capacity of the shorter plate must be compared with the rivet load transmitted to it. In this example, the upper cover plate transmits the rivet load of four rivets in single shear, or $4 \times 19.8 = 79.2 \text{ kN}$, which is less than its tear capacity of 111.2 kN. Hence, the load capacity of both cover plates becomes

$$P_c = 79.2 + 111.2 = 190.4 \text{ kN}$$

determined by rivet shear in the upper plate and by tension at row 3 in the lower plate. Thus, the safe load is the lowest of these several values = 169.0 kN, which is the rivet strength in shear.

$$\text{Efficiency} = \frac{\text{Safe load}}{\text{Strength of solid plate}} = \frac{169 \times 10^3}{(180 \times 10^3)(14 \times 10^{-3})(80 \times 10^6)} = 83.8\%$$

In this discussion, we have neglected friction and assumed that the rivets or bolts only act as pins in the structure or joint—in essence like spot welds spaced in the same way as the rivets or bolts are spaced.

7 FRICTION-TYPE CONNECTIONS

In friction-type connections, high-strength bolts of various grades are used and are tightened to high tensile stresses, thereby causing a large residual compression force between the plates. Tightening of the bolts to a predetermined initial tension is usually done using a calibrated torque wrench or by turn-of-the nut methods.

If done properly (as will be discussed later), transverse shear loads are now transferred by the friction between the plates and not by shear and the bearing of the bolt, as described in the previous sections. Heretofore, even though the bolts are not subject to shear, design codes, as a matter of convenience, specified an allowable shearing stress to be applied over the cross-sectional area of the bolt. Thus, friction-type joints were analyzed by the same procedures used for bearing-type joints, and the frictional forces that existed were taken as an extra factor of safety. In the American Society of Mechanical Engineers (ASME) code, the “allowable stresses” listed in several places are not intended to limit assembly stresses in the bolts. These allowables are intended to force flange designers to overdesign the joint to use more and/or larger bolts and thicker flange members than they might otherwise be inclined to use.

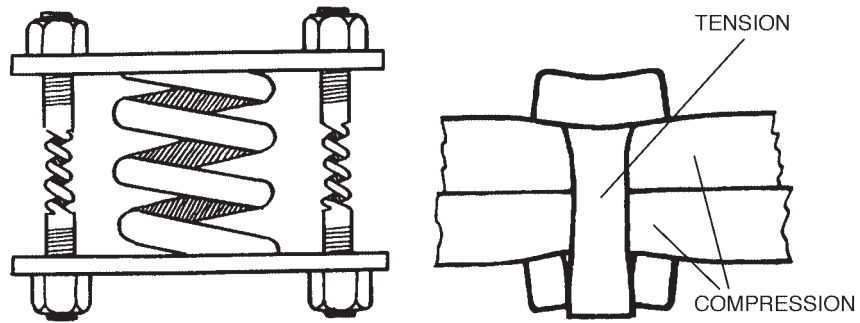


Figure 17 When analyzing the behavior of a bolted joint, pretend the members are a large spring being compressed (clamped) by a group of smaller springs (bolts). When tightened, these springs distort somewhat as shown but grossly exaggerated on the right.

Structurally, a bolt serves one of two purposes: It can act as a pin to keep two or more members from slipping relative to each other, or it can act as a heavy spring to clamp two or more pieces together.

In the vast majority of applications, the bolt is used as a clamp and, as such, it must be tightened properly. When we tighten a bolt by turning the head or the nut, we will stretch the bolt initially in the elastic region. More tightening past the elastic limit will cause the bolt to deform plastically. In either case, the bolt elongates and the plates deform in the opposite direction (equal compressive stresses in the materials being joined). In this way, you really have a spring system as shown (with substantial exaggeration) in Fig. 17.

The tensile stress introduced into the fastener during this initial tightening process results in a tension force within the fastener, which in turn creates the clamping force on the joint. This initial clamping force is called the *preload*. Preloading a fastener properly is a major challenge but is critical for fatigue in bolted systems.

When a bolt is loaded in a tensile testing machine, we generate a tension versus a change in length curve, as shown in Fig. 18. The initial straight-line portion of the elastic curve is called

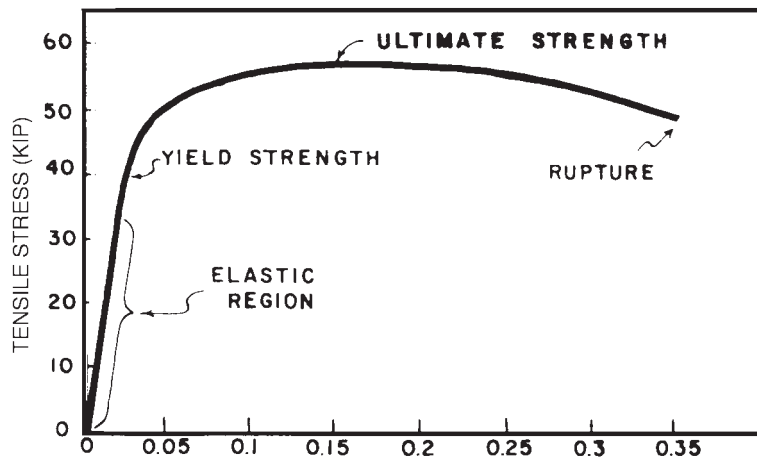


Figure 18 Engineering stress-strain curve (typical).

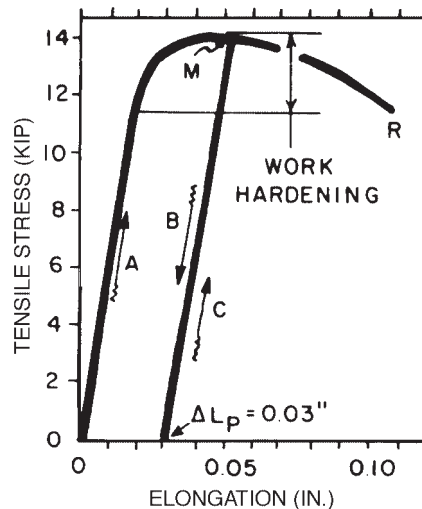


Figure 19 Elastic curve for a $\frac{3}{8}$ - 16 \times 4 socket-head cap screw loaded (A) to point M well past the yield strength and then unloaded (B) to give permanent deformation $L_p = 0.03$ in. If reloaded, it will follow path (C).

the *elastic region*. Loading and unloading a bolt within this range of tension never results in a permanent deformation of the bolt because elastic deformation is recoverable. The upper limit of this straight section of the curve ends at the *elastic limit*. Loading beyond or above this limit results in *plastic deformation* of the bolt, which is not recoverable; thus, the bolt has a permanent set (it is longer than it was originally even though the load is completely removed). At the *yield point*, the bolt has a specific amount of permanent plastic deformation, normally defined as 0.2 or 0.5% of the initial length. Permanent plastic deformation will increase up until the *ultimate tensile strength* (UTS), which is the maximum tension that can be carried in the bolt. The UTS is always greater than the yield stress, sometimes as much as twice yield. The final point on the curve is the *failure* or *rupture stress*, where the bolt breaks under the applied load.

If we load the bolt well into the plastic region of its curve and then remove the load, it will behave as shown in Fig. 19, returning to the zero load point along a line parallel to the original elastic line but offset by the amount of plastic strain the bolt has set.

On reloading the bolt below the previous load but above the original yield point, the behavior of the bolt will follow this new offset stress-strain line and the bolt will behave elastically well beyond the original load that caused plastic deformation in the first place. The difference between the original yield strength of the material and the new yield strength is a function of the work hardening that occurred by taking it past the original yield strength on the first cycle. By following the above procedure, we have made the bolt stronger, at least as far as static loads are concerned.

This is not wise practice, however, for more brittle materials can suffer a loss of strength by such treatments. Loss of strength in ASTM A490 bolts, because of repeated cycling past the yield (under water and wind loads), has been publicly cited as a contributing factor in the 1979 collapse of the roof on the Kemper Auditorium in Kansas City.

8 UPPER LIMITS ON CLAMPING FORCE

8.1 Design-Allowable Bolt Stress and Assembly Stress Limits

We need to follow the limits placed on bolt stresses by codes, company policies, and standard practices. Both structural steel and pressure vessel codes define maximum design allowable stresses for bolts. To distinguish between maximum design stress and the maximum stress that may be allowed in the fastener during assembly, we need to look at the design safety factor. These two will differ—that is, maximum design allowables will differ if a factor of safety is involved. For structural steels, bolts are frequently tightened well past the yield strength even though the design allowables are only 35–58% of yield. Pressure vessel bolts are commonly tightened to twice the design allowable. Aerospace, auto, and other industries may impose stringent limits on design stresses rather than on actual stresses to force the designer to use more or larger bolts.

8.2 Torsional Stress Factor

If the bolts are to be tightened by turning the nut or the head, they will experience a torsion stress as well as a tensile stress during assembly. If tightened to the yield stress, they will yield under this combination. If we plan to tighten to or near the yield stress, we must reduce the maximum tensile stresses allowed by a “torquing factor.” If using as received steel on steel bolts, then a reduction in the allowable tensile stress of 10% is reasonable. If the fasteners are to be lubricated, use 5%.

8.3 Other Design Issues

- *Flange Rotation.* Excessive bolt load can rotate raised face flanges so much that the ID of the gasket is unloaded, opening a leak path. The threat of rotation, therefore, can place an upper limit on planned or specified clamping forces.
- *Gasket Crush.* Excessive preload can so compress a gasket that it will not be able to recover when the internal pressure or a thermal cycle partially unloads it. Contact the gasket manufacturer for upper limits. Note that these will be a function of the service temperature.
- *Stress Cracking.* Stress cracking is encouraged by excessive tension in the bolts, particularly if service loads exceed 50% of the yield stress at least for low alloy quenched and tempered steels.
- *Combined Loads.* These loads include weight, inertial affects, thermal effects, pressure, shock, earthquake loading, and so on. Both static and dynamic loads must be estimated. Load intensifiers such as prying and eccentricity should be acknowledged if present. Joint diagrams can be used (see later section) to add external loads and preloading. The parts designed must be able to withstand *worst-case combinations of these pre- and service loads.*

Figure 20 shows the residual stresses in a group of 90 $2\frac{1}{4}$ –12 × 29 4330 studs that were tightened by stretching them 79% of their yield stress with a hydraulic tensioner. The studs and nuts were not new but had been tightened several times before these data were taken. Relaxation varied from 5 to 43% of the initial tension applied in these apparently identical studs. In many cases, similar scatter in relaxation also occurs after torquing.

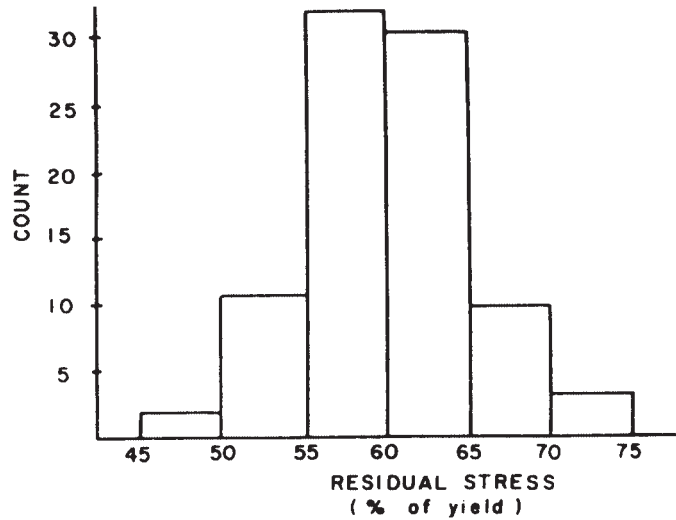


Figure 20 Residual stress as a percentage of yield strength, following removal of tension. Studs were all tensioned to 79% of yield. Torqued to 500 lb-ft.

Charts of this sort can be constructed on the basis of individual bolts or multibolt joints. Limits can be defined in terms of force, stress, yield instead of UTS, or even assembly torque.

9 THEORETICAL BEHAVIOR OF THE JOINT UNDER TENSILE LOADS

In this section, we examine the way a joint responds when exposed to the external loads it has been designed to support. This will be done by examining the elastic behavior of the joint. When we tighten a bolt on a flange, the bolt is placed in tension and it gets longer. The joint compresses in the vicinity of the bolt.

We need to plot separate elastic curves for the bolt and joint members by plotting the force in each of the two vertical axes and the deformation of each (elongation in the bolt and compression in the joint) on the horizontal axes. See Fig. 21.

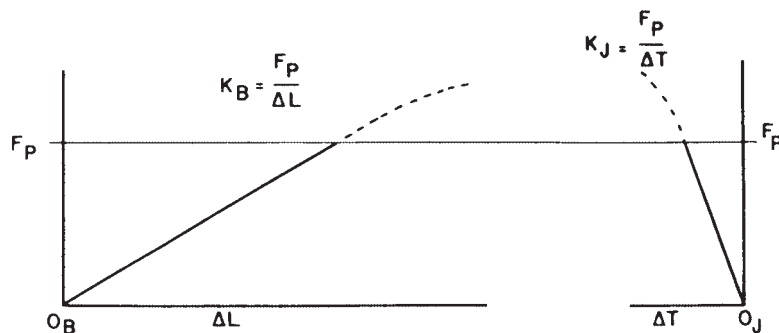


Figure 21 Elastic curves for bolt and joint members.

Three things should be noted:

1. Typically the slope (K_B) of the bolts elastic curve is only $1/3-1/5$ of the slope (K_J) of the joints elastic curve; i.e., the stiffness of the bolt is only $1/3-1/5$ that of the joint.
2. The clamping force exerted by the bolt on the joint is opposed by an equal and opposite force exerted by the joint members on the bolt. (The bolt wants to shrink back to its original length and the joint wants to expand to its original thickness.)
3. If we continue to tighten the bolt, it or the joint will ultimately start to yield plastically, as suggested by the dotted lines. In future diagrams, we will operate only in the elastic region of each curve.

Rotscher first demonstrated what is called a *joint diagram* (Fig. 22). In Fig. 22, the tensile force in the bolt is called the *preload* in the bolt and is equal and opposite to the compressive force in the joint. If we apply an additional tension force to the bolt, this added load partially relieves the load on the joint, allowing it (if enough load is applied) to return to its original thickness while the bolt gets longer. Note that the increase in the length of the bolt is equal to the increase in thickness (reduction in compression) in the joint. In other words, the *joint expands to follow the nut as the bolt lengthens*.

Because the stiffness of the bolt is only $1/3-1/5$ that of the joint, for an equal change in the strain, the change in load in the bolt must be only $1/3-1/5$ of the change in the load in the joint. This is shown in Fig. 23.

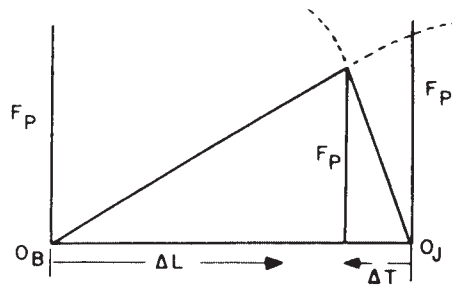


Figure 22 The elastic curves for bolt and joint can be combined to construct a joint diagram. O_B is the reference point for bolt length at zero stress. O_J is the reference point for joint thickness at zero stress.

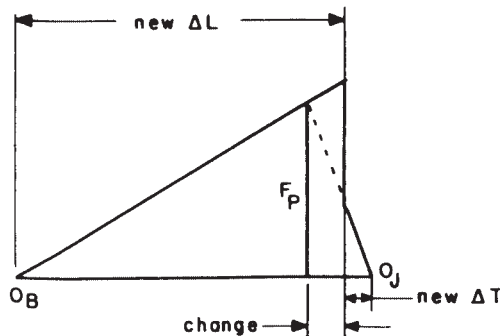


Figure 23 When an external tension load is applied, the bolt gets longer and joint compression is reduced. The change in deformation in the bolt equals the change in deformation in the joint.

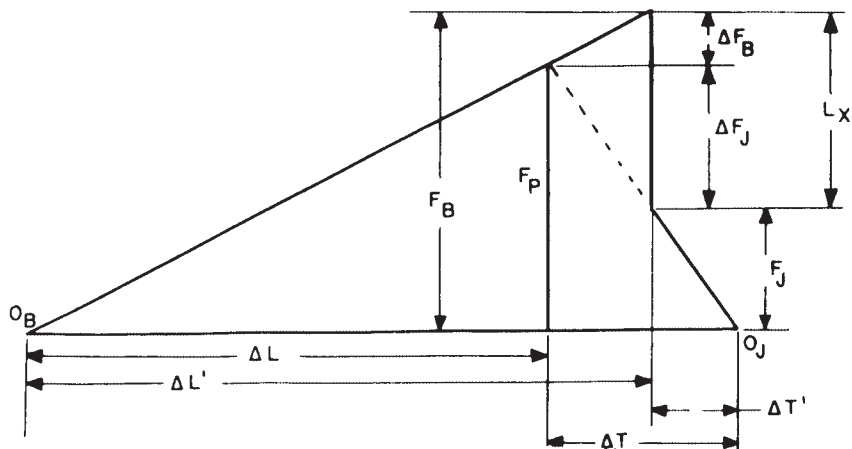


Figure 24 Summary diagram. F_P = initial preload; F_B = present bolt load; F_J = present joint load; L_X = external tension load applied to the bolt.

The external tension load (L_X) required to produce this change of force and strain in the bolt and joint members is equal to the increase in the force on the bolt (ΔF_B) plus the reduction of force in the joint (ΔF_J):

$$L_X = \Delta F_B + \Delta F_J$$

The above relationship is demonstrated in Fig. 24.

Any external tension load, no matter how small, will be partially absorbed in replacing the force in the bolt (ΔF_B) and partially absorbed in replacing the reduction of force that the joint originally exerted on the bolt (ΔF_J). The force of the joint on the bolt plus the external load equal the new total tension force in the bolt—which is greater than the previous total—but the change in bolt force is less than the external load applied to the bolt. See Fig. 24, which recaps all of this. This is extremely important because it is a way to move external loads around the bolt. In the case of alternating loads the fatigue life of the bolts can be greatly increased. That the bolt sees only a part of the external load, and that the amount it sees is dependent on the stiffness ratio, between the bolt and the joint, have many implications for joint design, joint failure, measurement of residual preloads, and so on.

We can change the joint stiffness between the bolt and the joint by making the bolt much stiffer (i.e., a bolt with a larger diameter). The new joint diagram resulting from this change is shown in Fig. 25. Note that the bolt now absorbs a larger percentage of the same external load.

9.1 Critical External Load Required to Overcome Preload

If we keep adding external load to the original joint, we reach a point where the joint members are fully unloaded, as in Fig. 26. This is the critical external load, which is not equal to the original preload in the bolt but is often equal to the preload for several reasons.

1. In many joints, the bolt has a low spring rate compared to the joint members. This is advantageous since the greatest percentage of external load is carried by the clamped members and not the bolt. Creation of a joint with a high ratio of K_J/K_B can be difficult. As an example, sheet-metal joints are extremely thin and therefore have a low value of K_J . There are several

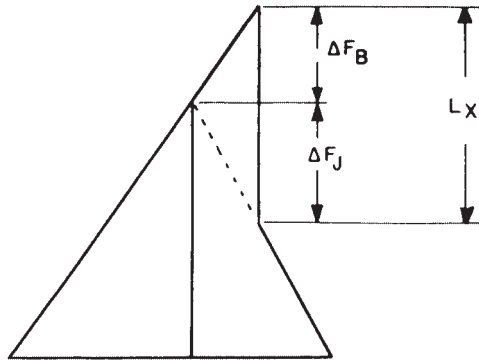


Figure 25 Joint diagram when stiffness of the bolt nearly equals that of joint.

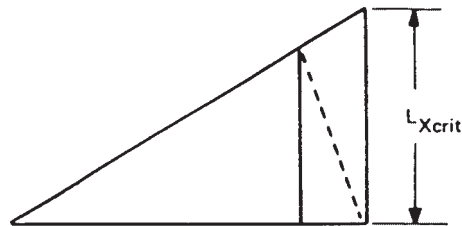


Figure 26 A critical external load (L_{Xcrit}) fully unloads the joint (but not the bolt). No load sharing exists.

strategies for lowering the stiffness of bolts in critical applications, including using hollow bolts as well as bolts with a reduced diameter in the middle.

2. Joints almost always relax after first tightening with relaxations of 10–20% of the initial preload being not uncommon. There are three main sources of relaxation in the joint. First the torsional load in the bolt is unstable and usually relaxes over time. Second, the nut is usually made of a more ductile material than the bolt and is designed to even out load among the threads. This is accomplished by a small movement of the threads relative to each other. Third, all of the mating surfaces that are compressed by the bolt contribute to relaxation. Entrapped particles as well as the high points of surface imperfections are crushed. If a bolt has $1/5$ the stiffness of the joint, then the critical external load to free the joint members is 20% greater than the residual preload in the bolt when the load is applied. Therefore, the difference between the critical external load and the present preload is just about equal and opposite to the loss in preload caused by bolt relaxation. In other words, the critical external load equals the original preload before bolt relaxation.

9.2 Very Large External Loads

Any additional external load we add beyond the critical point will all be absorbed by the bolt. Although it is usually ignored in joint calculations, there is another curve of which we should be aware. The compressive spring rate of many joint members is not a constant. A more accurate joint diagram would show this. See Fig. 27.

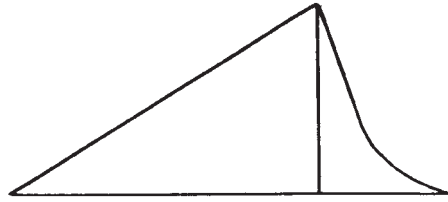


Figure 27 The spring rate of the joint is frequently nonlinear for small deflections.

For joint diagrams, as shown in Fig. 24, we can make the following calculations where

F_p = initial preload (lb, N)

L_x = external tension load (lb, N)

ΔF_B = change in load in bolt (lb, N)

ΔF_J = change in load in joint (lb, N)

ΔL , $\Delta L'$ = elongation of the bolt before and after application of the external load (in., mm)

ΔT , $\Delta T'$ = compression of joint members before and after application of the external load (in., mm)

$L_{x\text{critical}}$ = external load required to completely unload the joint (lb, N) (not shown in diagram)

The stiffness (spring constants) of the bolt and joint are defined as follows:

$$\text{for the bolt } K_B = \frac{F_p}{\Delta L}$$

$$\text{for the joint } K_J = \frac{F_p}{\Delta T}$$

by manipulation $\Delta F_B = \frac{K_B}{K_B + K_J} \times L_x$ until joint separation, after which

$$\Delta F_B = \Delta L_x \quad \text{and} \quad L_{x\text{critical}} = F_p \left\{ 1 + \frac{K_B}{K_J} \right\}$$

10 EVALUATION OF SLIP CHARACTERISTICS

A slip-resistant joint is one that has a low probability of slip at any time during the life of the structure. In this type of joint, the external applied load usually acts in a plane perpendicular to the bolt axis. The load is completely transmitted by frictional forces acting on the contact area of the plates fastened by the bolts. This frictional resistance is dependent on (1) the bolt preload and (2) the slip resistance of the fraying surfaces.

Slip-resistant joints are often used in connections subjected to stress reversals, severe stress fluctuations, or in any situation wherein slippage of the structure into a "bearing mode" would produce intolerable geometric changes. A slip load of a simple tension splice is given by

$$P_{\text{slip}} = k_s m \sum_{i=1}^n T_i$$

Where k_s = slip coefficient

m = number of slip planes

$\sum_{i=1}^n T_i$ = sum of the bolt tensions

Table 1 Summary of Slip Coefficients

Type of Steel	Treatment	Average	Standard Deviation	Number of Tests
A7, A36, A440	Clean mill scale	0.32	0.06	180
A7, A36, A440, Fe37, Fe.52	Clean mill scale	0.33	0.07	327
A588	Clean mill scale	0.23	0.03	31
Fe37	Grit blasted	0.49	0.07	167
A36, Fe37, Fe52	Grit blasted	0.51	0.09	186
A514	Grit blasted	0.33	0.04	17
A36, Fe37	Grit blasted, exposed	0.53	0.06	51
A36, Fe37, Fe52	Grit blasted, exposed	0.54	0.06	83
A7, A36, A514, A572	Sand blasted	0.52	0.09	106
A36, Fe37	Hot-dip galvanized	0.18	0.04	27
A7, A36	Semipolished	0.28	0.04	12
A36	Vinyl wash	0.28	0.02	15
	Cold zinc plated	0.30	—	3
	Metallized	0.48	—	2
	Galvanized and sand blasted	0.34	—	1
	Sand blasted treated with linseed oil (exposed)	0.26	0.01	3
	Red lead paint	0.06	—	6

If the bolt tension is equal in all bolts, then

$$P_{\text{slip}} = k_s n m T_i$$

where n is the number of bolts in the joint.

The slip coefficient K_s varies from joint to joint, depending on the type of steel, different surface treatments, and different surface conditions, and along with the clamping force T_i shows considerable variation from its mean value. The slip coefficient K_s can only be determined experimentally, but some values are now available, as shown in Table 1.

11 TURN-OF-NUT METHOD OF BOLT TIGHTENING

To overcome the variability of torque control efforts a more reliable tightening procedure is called the turn-of-nut method. (This is a strain-control method.) Initially it was believed that one turn from the snug position was the key, but because of out of flatness, thread imperfections, and dirt accumulation, it was difficult to determine the hand-tight position (the starting point—from the snug position). Current practice is as follows: Run the nut up to a snug position using an impact wrench rather than the finger-tight condition (elongations are still within the elastic range). From the snug position, turn the nut in accordance with Table 2, provided by the RCSC specification.

Nut rotation is relative to bolt, regardless of the element (nut or bolt) being turned. For bolts installed by $2/3$ turn and less, the tolerance should be $\pm 30^\circ$; for bolts installed by $2/3$ turn and more, the tolerance should be $\pm 45^\circ$. All material within the grip of the bolt must be steel.

No research work has been performed by the council to establish the turn-of-nut procedure when bolt length exceeds 12 diameters. Therefore, the required rotation must be determined by actual tests in a suitable tension device simulating the actual conditions.

When bolts pass through a sloping interface greater than 1:20, a beveled washer is required to compensate for the lack of parallelism. As noted in Table 2, bolts require additional nut rotation to ensure that tightening will achieve the required minimum preload.

Table 2 Nut Rotation from Snug-Tight Condition

Bolt Length (as measured from underside of head to extreme end of point)	Both Faces Normal to Bolt Axis	One Face Normal to Bolt Axis and Other Face Sloped Not More Than 1:20 (bevel washer not used)	Both Faces Sloped Not More Than 1:20 from Normal to Bolt Axis (bevel washers not used)
Up to and including 4 diameters	1/3 turn	1/2 turn	2/3 turn
Over 4 diameters but not exceeding 8 diameters	1/2 turn	2/3 turn	5/6 turn
Over 8 diameters but not exceeding 12 diameters	2/3 turn	5/6 turn	1 turn

12 TORQUE AND TURN TOGETHER

Measuring of torque and turn at the same time can improve our control over preload. The final variation in preload in a large number of bolts is closer to $\pm 5\%$ than the 25–30% if we used torque or turn control alone. For this reason the torque–turn method is widely used today, especially in structural steel applications.

In this procedure, the nut is first snugged with a torque that is expected to stretch the fastener to a minimum of 75% of its ultimate strength. The nut is then turned (half a turn) or the like, which stretches the bolt well past its yield point. See Fig. 28.

This torque–turn method cannot be used on brittle bolts, but only on ductile bolts having long plastic regions. Therefore, it is limited to A325 fasteners used in structural steel work. Furthermore, it should never be used unless you can predict the working loads that the bolt will see in service. Anything that loads the bolts above the original tension will create additional plastic deformation in the bolt. If the overloads are high enough, the bolt will break.

A number of knowledgeable companies have developed manual torque–turn procedures that they call “turn of the nut” but that do not involve tightening the fasteners past the yield point. Experience shows that some of these systems provide additional accuracy over turn or torque alone.

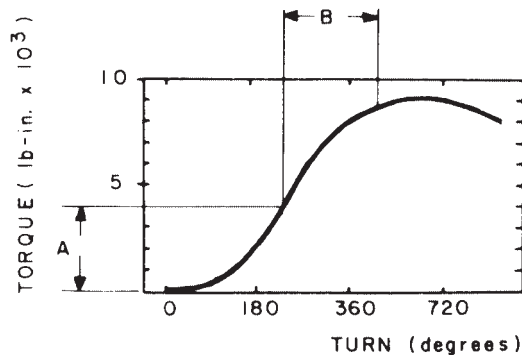


Figure 28 In turn-of-nut techniques, the nut is first tightened with an approximate torque (A) and then further tightened with a measured turn (B).

Other methods have also been developed to control the amount of tension produced in bolts during assembly, namely *stretch* and *tension control*.² All of these methods have drawbacks and limitations, but each is good enough for many applications. However, in more and more applications, we need to find a better way to control bolt tension and/or clamping forces. Fortunately, that better way is emerging, namely *ultrasonic measurement of bolt stretch or tension*.

13 ULTRASONIC MEASUREMENT OF BOLT STRETCH OR TENSION

Ultrasonic techniques, while not in common use, allow us to get past dozens of the variables that affect the results we achieve with torque and/or torque and turn control.

The basic concepts are simple. The two most common systems are *pulse-echo* and *transit time* instruments. In both, a small acoustic transducer is placed against one end of the bolt being tested. See Fig. 29. An electronic instrument delivers a voltage pulse to the transducer, which emits a very brief burst of ultrasound that passes down the bolt, echoes off the far end, and returns to the transducer. An electronic instrument measures precisely the amount of time required for the sound to make its round trip in the bolt.

As the bolt is tightened, the amount of time required for the round trip increases for two reasons:

1. The bolt stretches as it is tightened, so the path length increases.
2. The average velocity of sound within the bolt decreases because the average stress level in the bolt has increased.

Both of these changes are linear functions of the preload in the fastener, so that the total change in transit time is also a linear function of preload.

The instrument is designed to measure the change in transit time that occurs during tightening and to report the results as

1. A change in length of the fastener
2. A change in the stress level within the threaded region of the fastener
3. A change in tension within the fastener

Using such an instrument is relatively easy. A drop of coupling fluid is placed on one end of the fastener to reduce the acoustic impedance between the transducer and the bolt. The transducer

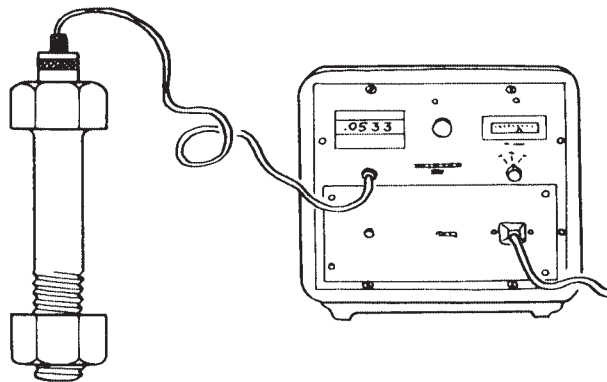


Figure 29 An acoustic transducer is held against one end of the fastener to measure the fastener's change in length as it is tightened.

is placed on the puddle of fluid and held against the bolt, mechanically or magnetically. The instrument is zeroed for this particular bolt (because each bolt will have a slightly different acoustic length). If you wish to measure residual preload, or relaxation, or external loads at some later date, you record the length of the fastener at zero load at this time. Next the bolt is tightened. If the transducer can remain in place during tightening, the instrument will show you the buildup of stretch or tension in the bolt. If it must be removed, it is placed on the bolt after tightening to show the results achieved by torque, turn, or tension.

If, at some later date, you wish to measure the present tension, you dial in the original length of *that* bolt into the instrument and place the transducer back on the bolt. The instrument will then show you the difference in length or stress that now exists in the bolt.

Ultrasonic equipment is used primarily in applications involving relatively few bolts in critically important joints or quality control audits. Operator training in the use of this equipment is necessary and is a low-cost alternative to strain-gauged bolts in all sorts of studies. The author knows of at least one application where ultrasonic measurement is built into a production work cell. It is used to tighten the crankshaft main bearings to very precise values.

These instruments are new to the field, so you must be certain to find out from the manufacturers exactly what the equipment will or will not do as well as precise information needed for use or equipment calibration. Training is essential not only for the person ordering the equipment but for all who will use it in the field or laboratory. Proper calibration is essential. If the equipment can only measure transit time, you must tell it how to interpret transit time data *for your application*.

14 FATIGUE FAILURE AND DESIGN FOR CYCLICAL TENSION LOADS

A fastener subjected to repeated cyclical tension loads can suddenly break. These failures are generally catastrophic in kind, even if the loads are well below the yield strength of the material.

Three essential conditions are necessary for a fatigue failure: cyclical tensile loads; stress levels above the endurance limit of the material; and a stress concentration region (such as a sharp corner, a hole, a surface scratch or other mark on the surface of the part, corrosion pits, an inclusion and/or a flaw in the material). Essentially no part is completely free of these types of defects unless great care has been taken to remove them.

The sequence of events leading up to a fatigue failure is as follows:

1. Crack initiation begins after about 90% of the total fatigue life (total number of cycles) has occurred. This crack always starts on the surface of the part.
2. The crack begins to grow with each half-cycle of tension stress, leaving beach marks on the part.
3. Growth of the crack continues until the now-reduced cross section is unable to support the load, at which time the part fails catastrophically.

A bolt is a very poor shape for good fatigue resistance. Although the average stress levels in the body may be well below the endurance limit, stress levels in the stress concentration points, such as thread roots, head to body fillets, and so on can be well over the endurance limit. One thing we can do to reduce or eliminate a fatigue problem is to attempt to overcome one or more of the three essential conditions without which failure would not occur. In general, most of the steps are intended to reduce stress levels, reduce stress concentrations, and/or reduce the load excursions seen by the bolt. The following are additional suggestions for reducing the chance of bolt fatigue.

Rolled Threads

Rolling provides a smoother thread finish than cutting and thus lowers the stress concentrations found at the root of the thread. In addition to overcoming the notch effect of cut threads, rolling

induces compressive stresses on the surface rolled. This compressive “preload” must be overcome by tension forces before the roots will be in net tension. A given tension load on the bolt, therefore, will result in a smaller tension excursion at this critical point. Rolling the threads is best done after heat treating the bolt, but it is more difficult. Rolling before heat treatment is possible on larger-diameter bolts.

Fillets

Use bolts with generous fillets between the head and the shank. An elliptical fillet is better than a circular one and the larger the radius the better. Prestressing the fillet is wise (akin to thread rolling).

Perpendicularity

If the face of the nut, the underside of the bolt head, and/or joint surfaces are not perpendicular to thread axes and bolt holes, the fatigue life of the bolt can be seriously affected. For example, a 2° error reduces the fatigue life by 79%.³

Overlapping Stress Concentrations

Thread run-out should not coincide with a joint interface (where shear loads exist) and there should be at least two full bolt threads above and below the nut because bolts normally see stress concentrations at (1) thread run-out and (2) first threads to engage the nut, and head-to-shank fillets.

Thread Run-Out

The run-out of the thread should be gradual rather than abrupt. Some people suggest a maximum of 15° to minimize stress concentrations.

Thread Stress Distribution

Most of the tension in a conventional bolt is supported by the first two or three nut threads. Anything that increases the number of active threads will reduce the stress concentration and increase the fatigue life. Some of the possibilities are

1. Using so-called tension nuts, which create nearly uniform stress in all threads.
2. Modifying the nut pitch so that it is slightly different than the pitch of the bolt, i.e., thread of nut 11.85 threads/in. used with a bolt having 12 threads/in.
3. Using a nut slightly softer than the bolt (this is the usual case); however, select still softer nuts if you can stand the loss in proof load capability.
4. Using a jam nut, which improves thread stress distribution by preloading the threads in a direction opposite to that of the final load.
5. Tapering the threads slightly. This can distribute the stresses more uniformly and increase the fatigue life. The taper is 15°.

Bending

Reduce bending by using a spherical washer because nut angularity hurts fatigue life.

Corrosion

Anything that can be done to reduce corrosion will reduce the possibilities of crack initiation and/or crack growth and will extend fatigue life. Corrosion can be more rapid at points of high stress concentration, which is also the point where fatigue failure is most prevalent. Fatigue and corrosion aid each other and it is difficult to tell which mechanism initiated or resulted in a failure.

Surface Conditions

Any surface treatment that reduces the number and size of incipient cracks will improve fatigue life significantly, so that polishing of the surface will greatly improve the fatigue life of a part. This is particularly important for punched or drilled holes, which can be improved by reaming and expanding to put the surface in residual compression. Shot peening of bolts or any surface smooths out sharp discontinuities and puts the surface in residual compression. Handling of bolts in such a way as not to ding one against the other is also important.

Reduce Load Excursions

It is necessary to identify the maximum safe preload that your joint can stand by estimating fastener strength, joint strength, and external loads. Also do whatever is required to minimize the bolt-to-joint stiffness ratio so that most of the excursion and external load will be seen by the joint and not the bolt. Use long, thin bolts even if it means using more bolts. Eliminate gaskets and/or use stiffer gaskets.

While there are methods available for estimating the endurance limit of a bolt, it is best to base your calculations on actual fatigue tests of the products you are going to use or your own experience with those products.

For the design criteria for fatigue loading of slip resistant joints, see Refs. 2 and 3.

REFERENCES

1. D. Whitney, *Mechanical Assembly*, Oxford University Press, New York, 2004.
2. J. H. Bickford, *An Introduction to the Design and Behavior of Bolted Joints*, 2nd ed., Marcel Dekker, New York, 1990.
3. G. L. Kulak, J. W. Fisher, and J. H. A. Struik, *Guide to Design Criteria for Bolted and Riveted Joints*, Wiley, New York, 1987.
4. *SPS Fastener Facts*, Standard Pressed Steel Co., Jenkintown, PA, Section IV-C-4.
5. G. Linnert, *Welding, Metallurgy, Carbon and Alloy Steels*, Vol. 4, American Welding Society, Miami, FL, 1994, Chap. 7.
6. N. Yurioka, "Weldability of Modern High Strength Steels," in *First US–Japan Symposium on Advances in Welding Metallurgy*, American Welding Society, Miami, FL, 1990, pp. 79–100.

CHAPTER 9

SEAL TECHNOLOGY

Bruce M. Steinetz
NASA Glenn Research Center at Lewis Field
Cleveland, Ohio

1 INTRODUCTION	283		
2 STATIC SEALS	283		
2.1 Gaskets	283		
2.2 O-Rings	289		
2.3 Packings and Braided Rope Seals	292		
3 DYNAMIC SEALS	296		
3.1 Initial Seal Selection	296		
3.2 Mechanical Face Seals	296		
3.3 Emission Concerns	301		
		3.4 Noncontacting Seals for High-Speed/Aerospace Applications	304
		3.5 Labyrinth Seals	308
		3.6 Honeycomb Seals	312
		3.7 Brush Seals	313
		3.8 Ongoing Developments	319
		REFERENCES	319
		BIBLIOGRAPHY	323

1 INTRODUCTION

Seals are required to fulfill critical needs in meeting the ever-increasing system performance requirements of modern machinery. Approaching a seal design, one has a wide range of available seal choices. This chapter aids the practicing engineer in making an initial seal selection and provides current reference material to aid in the final design and application.

This chapter provides design insight and application for both static and dynamic seals. Static seals reviewed include gaskets, O-rings, and selected packings. Dynamic seals reviewed include mechanical face, labyrinth, honeycomb, and brush seals. For each of these seals, typical configurations, materials, and applications are covered. Where applicable, seal flow models are presented.

2 STATIC SEALS

2.1 Gaskets

Gaskets are used to effect a seal between two mating surfaces subjected to differential pressures. Gasket types and materials are limited only by one's imagination. Table 1 lists some common gasket materials and Table 2 lists common elastomer properties.¹ The following gasket characteristics are considered important for good sealing performance.² Selecting the gasket material that has the best balance of the following properties will result in the best practical gasket design.

- Chemical compatibility
- Heat resistance
- Compressibility

Table 1 Common Gasket Materials, Gasket Factors (m), and Minimum Design Seating Stress (y)


















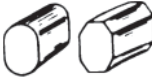
Gasket Material	Gasket Factor m	Min. Design Seating Stress y , psi	Sketches
Self-energizing types (O-rings, metallic, elastomer, other gasket types considered as self-sealing)	0	0	...
<i>Elastomers without fabric or high percent of asbestos fiber:</i>			
Below 75A Shore durometer	0.50	0	
75A or higher Shore durometer	1.00	200	
<i>Asbestos with suitable binder for operating conditions:</i>			
1/8 in. thick	2.00	1,600	
1/16 in. thick	2.75	3,700	
1/32 in. thick	3.50	6,500	
Elastomers with cotton fabric insertion	1.25	400	
<i>Elastomers with asbestos fabric insertion (with or without wire reinforcement):</i>			
3-ply	2.25	2,200	
2-ply	2.50	2,900	
1-ply	2.75	3,700	
Vegetable fiber	1.75	1,100	
<i>Spiral-wound metal, asbestos filled:</i>			
Carbon	2.50	10,000	
Stainless, Monel, and nickel-based alloys	3.00	10,000	
<i>Corrugated metal, asbestos inserted, or corrugated metal, jacketed asbestos filled:</i>			
Soft aluminum	2.50	2,900	
Soft copper or brass	2.75	3,700	
Iron or soft steel	3.00	4,500	
Monel or 4–6% chrome	3.25	5,500	
Stainless steels and nickel-based alloys	3.50	6,500	
<i>Corrugated metal:</i>	2.75	3,700	
Soft aluminum			
Soft copper or brass	3.00	4,500	
Iron or soft steel	3.25	5,500	
Monel or 4–6% chrome	3.50	6,500	
Stainless steels and nickel-based alloys	3.75	7,600	
<i>Flat metal, jacketed asbestos filled:</i>			
Soft aluminum	3.25	5,500	
Soft copper or brass	3.50	6,500	
Iron or soft steel	3.75	7,600	
Monel	3.50	8,000	
4–6% chrome	3.75	9,000	
Stainless steels and nickel-based alloys	3.75	9,000	

Table 1 (continued)

Gasket Material	Gasket Factor m	Min. Design Seating Stress y , psi	Sketches
<i>Grooved metal:</i>			
Soft aluminum	3.25	5,500	
Soft copper or brass	3.50	6,500	
Iron or soft steel	3.75	7,600	
Monel or 4–6% chrome	3.75	9,000	
Stainless steels and nickel-based alloys	4.25	10,100	
<i>Solid flat metal:</i>			
Soft aluminum	4.00	8,800	
Soft copper or brass	4.75	13,000	
Iron or soft steel	5.50	18,000	
Monel or 4–6% chrome	6.00	21,800	
Stainless steels and nickel-based alloys	6.50	26,000	
<i>Ring joint:</i>			
Iron or soft steel	5.50	18,000	
Monel or 4–6% chrome	6.00	21,800	
Stainless steels and nickel-based alloys	6.50	26,000	

Source: Table 2-5.1, ASME Code for Pressure Vessels, 1995.

- Microconformability (asperity sealing)
- Recovery
- Creep relaxation
- Erosion resistance
- Compressive strength (crush resistance)
- Tensile strength (blowout resistance)
- Shear strength (flange shearing movement)
- Removal, or “Z,” strength
- Antistick
- Heat conductivity
- Acoustic isolation
- Dimensional stability

Nonmetallic Gaskets. Most *nonmetallic gaskets* consist of a fibrous base held together with some form of an elastomeric binder. A gasket is formulated to provide the best load-bearing properties while being compatible with the fluid being sealed.

Nonmetallic gaskets are often reinforced to improve torque retention and blowout resistance for more severe service requirements. Some types of reinforcements include perforated cores, solid cores, perforated skins, and solid skins, each suited for specific applications. After a gasket material has been reinforced by either material additions or laminating, manufacturers can emboss the gasket raising a sealing lip, which increases localized pressures, thereby increasing sealability.

Metallic Gaskets. *Metallic gaskets* are generally used where either the joint temperature or load is extreme or in applications where the joint might be exposed to particularly caustic chemicals.

Table 2 The Most Important Elastomers and Their Properties

Elastomer	Composition	Working temperature range, °C	Tensile strength, bars	Elongation, %	Hardness, °Shore	Water	Steam	Hydraulic fluids, non-flammable (ester based)	Mineral fats and oils	Vegetable and animal fats and oils	Ozone	Hydrocarbons									
												Aliphatic	Aromatic	Halogenated	Alcohols	Ketones	Esters	Dilute acids	Concentrated acids	Dilute alkalis	Concentrated alkalis
Natural rubber	Rubber, K. W. Coil Refining-type polymerisate	-30-120	50-280	1000	30-98	x	x	—	—	—	—	x	x	—	0	—	x	0	x	0	x
SBR	Butadiene-styrene copolymer	-30-130	50-240	700	40-95	x	x	—	—	—	0	x	x	—	x	0	x	x	x	x	x
Nitrile N	Butadiene-acrylonitrile copolymer	-30-130	50-240	700	40-95	x	0	—	x	x	0	x	—	—	0	—	0	0	x	0	x
Neoprene	Chlorinated-butadiene polymerisate	-40-140	50-270	800	40-95	x	x	—	0	0	x	0	—	—	x	0	x	x	x	x	x
Butyl	Isobutylene-isoprene copolymer	-50-150	40-170	900	40-90	x	x	0	—	0	x	0	—	—	x	0	x	0	x	x	x
Hypalon	Chlorosulfonated polyethylene	-40-140	40-200	600	40-95	x	0	—	—	0	x	—	—	—	x	0	x	0	x	x	x
Silicone rubber	Polycondensates of dialkylsiloxanes	-100-200	20-80	500	40-80	0	—	—	0	x	x	0	—	—	x	0	x	0	x	0	0
Thiokol	Alkylpolysulfide	-40-80	10-60	200	65-80	x	—	x	x	x	x	0	0	x	0	x	x	0	x	x	x
Polyacrylic	Polyacrylate	-30-120	20-70	700	70-85	0	—	x	x	x	x	0	0	—	—	—	0	—	—	—	0
Vulcollan	Polyurethane	-30-80	200-320	600	70-95	0	—	—	x	x	x	—	—	—	—	—	—	—	—	—	—
Adiprene	Polyurethane	-40-120	80-300	700	70-95	x	0	—	x	x	x	—	—	—	—	—	—	—	—	—	—
Kel-F	Copolymer of chlorotriethylene and vinylidene fluoride	-50-180	30-120	700	60-90	x	x	—	—	0	x	0	—	—	x	x	x	x	x	x	x
Viton	Vinylidene fluoride-hexafluoropropylene copolymer	-60-200	80-160	300	60-95	x	0	0	x	x	x	0	x	—	x	0	0	0	—	—	x
PTFE	Polytetrafluoroethylene	-200-280	140-310	200	55D	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
EPR	Ethylene-propylene	-55-200	50-160	400	70-95	x	x	x	—	—	x	—	—	—	x	0	x	0	x	x	x
FSR	Fluorosilicone rubber	-60-230	55-85	400	40-80	0	0	0	x	x	x	0	x	—	x	0	x	—	x	0	x

Note: From Ref. 1, x, stable; 0, stable under certain conditions; —, unstable.

A good seal capable of withstanding very high temperature is possible if the joint is designed to yield locally over a narrow location with application of bolt load. Some of the most common metallic gaskets range from soft varieties, such as copper, aluminum, brass, and nickel, to highly alloyed steels. Noble metals, such as platinum, silver, and gold, also have been used in difficult locations.

Metallic gaskets are available in both standard and custom designs. Since there is such a wide variety of designs and materials used, it is recommended that the reader directly contact metallic gasket suppliers for design and sealing information.

Required Bolt Load: ASME Method

The American Society of Mechanical Engineers (ASME) Code for Pressure Vessels, Section VIII, Division 1, Appendix 2, is the most commonly used design method for gasketed joints where important joint properties, including flange thickness and bolt size and pattern, are specified. An integral part of the ASME Code revolves around two gasket factors:

1. An m factor, often called the gasket maintenance factor, is associated with the hydrostatic end force and the operation of the joint.
2. The y factor is a rough measure of the minimum seating stress associated with a particular gasket material. The y factor pertains only to the initial assembly of the joint.

The ASME Code makes use of two basic equations to calculate bolt load, with the larger calculated load being used for design:

$$W_{m1} = H + H_p = \frac{\pi}{4} G^2 P + 2\pi b G m P$$

$$W_{m2} = H_y = \pi b G y$$

where W_{m1} = minimum required bolt load from maximum operating or working conditions, lb

W_{m2} = minimum required initial bolt load for gasket seating (atmospheric temperature conditions) without internal pressure, lb

H = total hydrostatic end force, lb $[(\pi/4)G^2P]$

H_p = total joint–contact–surface compression load, lb

H_y = total joint–contact–surface seating load, lb

G = diameter at location of gasket load reaction; generally defined as follows:

when $b_0 < 1/4$ in., G = mean diameter of gasket contact face, in.; when $b_0 > 1/4$ in., G = outside diameter of gasket contact face less $2b$, in.

P = maximum internal design pressure, psi

b = effective gasket or joint–contact–surface seating width, in; $= b_0$ when $b_0 \leq 1/4$ in., $= 0.5\sqrt{b_0}$ when $b_0 > 1/4$ in. (see also ASME Table 2-5.2)

$2b$ = effective gasket or joint–contact–surface pressure width, in.

b_0 = basic gasket seating width

m = gasket factor per ASME Table 2-5.1 (repeated here as Table 1)

y = gasket or joint–contact–surface unit seating load, per ASME Table 2-5.1 (repeated here as Table 1), psi

The factor m provides a margin of safety to be applied when the hydrostatic end force becomes a determining factor. Unfortunately, this value is difficult to obtain experimentally since it is not a constant. The equation for W_{m2} assumes that a certain unit stress is required on a gasket to make it conform to the sealing surfaces and be effective. The second empirical constant y represents the gasket yield–stress value and is very difficult to obtain experimentally.

Practical Considerations

Flange Surfaces. Preparing the flange surfaces is paramount for effecting a good gasket seal. Surface finish affects the degree of sealability. The rougher the surface, the more bolt load required to provide an adequate seal. Extremely smooth finishes can cause problems for high operating pressures, as lower frictional resistance leads to a higher tendency for blowout. Surface finish lay is important in certain applications to mitigate leakage. Orienting finish marks transverse to the normal leakage path will generally improve sealability.

Flange Thickness. Flange thickness must also be sized correctly to transmit bolt clamping load to the area between the bolts. Maintaining seal loads at the midpoint between the bolts must be kept constantly in mind. Adequate thickness is also required to minimize the bowing of the flange. If the flange is too thin, the bowing will become excessive and no bolt load will be carried to the midpoint, preventing sealing.

Bolt Pattern. Bolt pattern and frequency are critical in effecting a good seal. The best bolt clamping pattern is invariably a combination of the maximum practical number of bolts, optimum spacing, and positioning.

One can envision the bolt loading pattern as a series of straight lines drawn from bolt to adjacent bolt until the circuit is completed. If the sealing areas lie on either side of this pattern, it will likely be a potential leakage location. Figure 1 shows an example of the various conditions.² If bolts cannot be easily repositioned on a problematic flange, Fig. 2 illustrates techniques to improve gasket effectiveness through reducing gasket face width where bolt load is minimum. Note that gasket width is retained in the vicinity of the bolt to support local bolt loads and minimize gasket tearing.

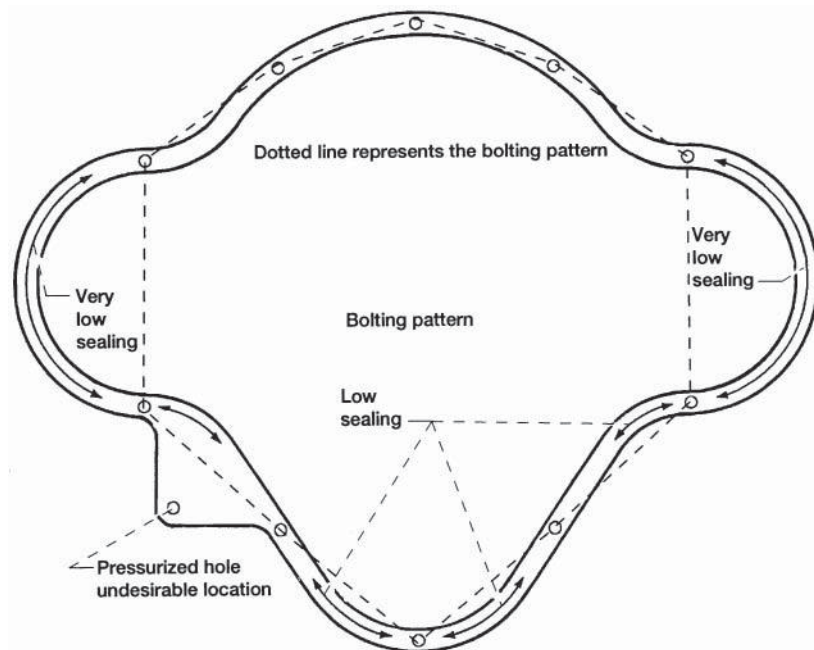


Figure 1 Bolting pattern indicating poor sealing areas. (Source: From Ref. 2.)

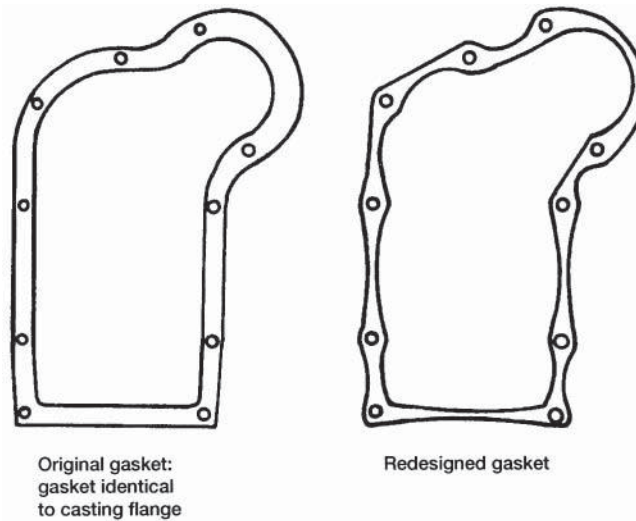


Figure 2 Original versus redesigned gasket for improved sealing. (Source: From Ref. 2.)

Gasket Thickness and Compressibility. Gasket thickness and compressibility must be matched to the rigidity, roughness, and unevenness of the mating flanges. An effective gasket seal is achieved only if the stress level imposed on the gasket at installation is adequate for the specific gasket and joint requirements.

Gaskets made of compressible materials should be as thin as possible. Adequate gasket thickness is required to seal and conform to the unevenness of the mating flanges, including surface finish, flange flatness, and flange warpage during use. A gasket that is too thick can compromise the seal during pressurization cycles and is more likely to exhibit creep relaxation over time.

Elevated-Temperature Service. Use of gaskets at elevated temperatures results in some additional challenges. The Pressure Vessel Research Council of the Welding Research Council has published several bulletins in this area (see Refs. 3 and 4).

2.2 O-Rings

O-ring seals are perhaps one of the most common forms of seals. Following relatively straightforward design guidelines, a designer can be confident of a high-quality seal over a wide range of operating conditions. This section provides useful insight to designers approaching an O-ring seal design, including the basic sealing mechanism, preload, temperature effects, common materials, and chemical compatibility with a range of working fluids. The reader is directed to manufacturer's design manuals for detailed information on the final selection and specification.⁵

Basic Sealing Mechanism

O-rings are compressed between the two mating surfaces and are retained in a seal gland. The initial compression provides initial sealing critical to successful sealing. Upon increase of the pressure differential across the seal, the seal is forced to flow to the lower pressure side of the gland (see Fig. 3). As the seal moves, it gains greater area and force of sealing contact.

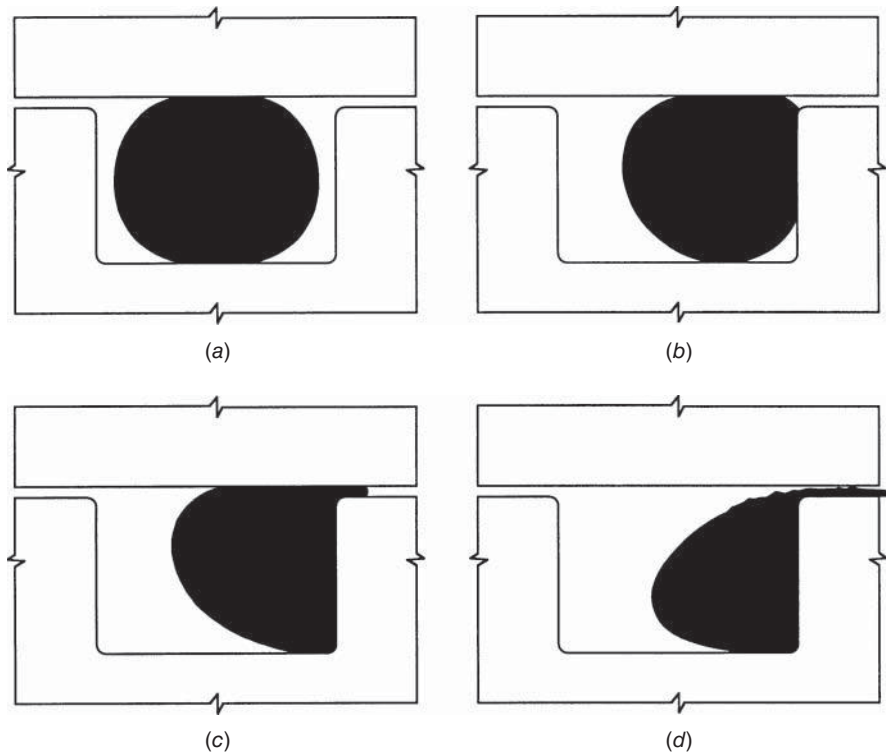


Figure 3 Basic O-ring sealing mechanism: (a) O-ring installed; (b) O-ring under pressure; (c) O-ring extruding; (d) O-ring failure. (Source: From Ref. 5, Reprinted courtesy of Parker Hannifin. Year of copyright is first year indicated herein. All rights reserved.)

At the pressure limit of the seal, the O-ring just begins to extrude into the gap between the inner and outer member of the gap. If this pressure limit is exceeded, the O-ring will fail by extruding into the gap. The shear strength of the seal material is no longer sufficient to resist flow and the seal material extrudes (flows) out of the open passage. Backup rings are used to prevent seal extrusion for high-pressure static and for dynamic applications.

Preload

The tendency of an O-ring to return to its original shape after the cross section is compressed is the basic reason why O-rings make such excellent seals. The maximum linear compression suggested by manufacturers is 30% for static applications and 16% for dynamic seals (up to 25% for small cross-sectional diameters). Compression less than these values is acceptable, within reason, if assembly problems are an issue. Manufacturers recommend⁵ a minimum amount of initial linear compression to overcome the compression set that O-rings exhibit.

O-ring compression force depends principally on the hardness of the O-ring, its cross-sectional dimension, and the amount of compression. Figure 4 illustrates the range of compressive force per linear inch of seal for typical linear percent compressions (0.139 in. cross-sectional diameter) and compound hardness (Shore A hardness scale). Softer compounds provide better sealing ability, as the rubber flows more easily into the grooves. Harder compounds are specified for high pressures, to limit chance of extruding into the groove, and to improve wear life for dynamic service. For most applications, compounds having a type A durometer hardness of 70–80 are the most suitable compromise.⁵

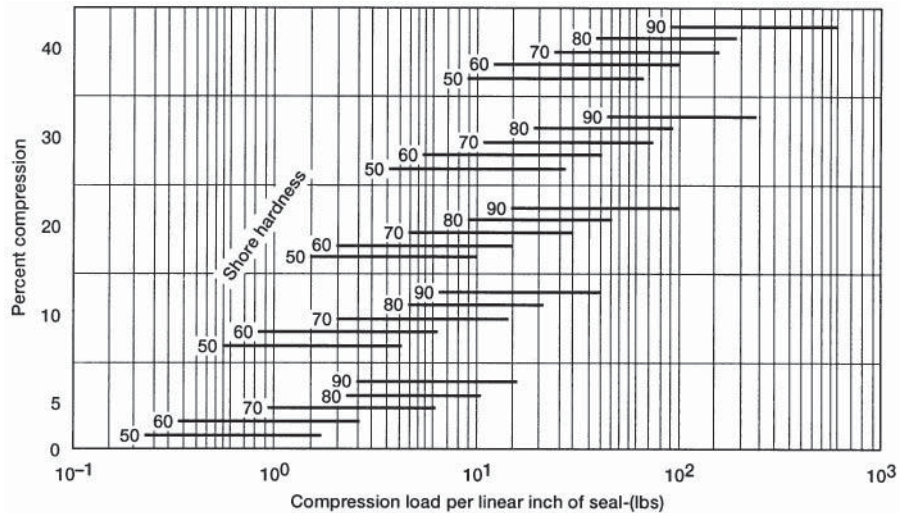


Figure 4 Effect of percent compression and material Shore hardness on seal compression load, 0.139 in. cross section. (Source: From Ref. 5, Reprinted courtesy of Parker Hannifin. Year of copyright is first year indicated herein. All rights reserved.)

Thermal Effects

O-ring seals respond to temperature changes. Therefore, it is critical to ensure the correct material and hardness are selected for the application. High temperatures soften compounds. This softening can negatively affect the seal's extrusion resistance at temperature. Over long periods of time at high temperature, chemical changes occur. These generally cause an increase in hardness, along with volume and compression set changes.

O-ring compounds harden and contract at cold temperatures. These effects can both lead to a loss of seal if initial compression is not set properly. Because the compound is harder, it does not flow into the mating surface irregularities as well. Just as important, the more common O-ring materials have a coefficient of thermal expansion (CTE) 10 times greater than that of steel (i.e., nitrile CTE is 6.2×10^{-5} per °F).

Groove dimensions must be sized correctly to account for this dimensional change. Manufacturers' design charts⁵ are devised such that proper O-ring sealing is ensured for the temperature ranges for standard elastomeric materials. However, the designer may want to modify gland dimensions for a given application that experiences only high or low temperatures in order to maintain a particular squeeze on the O-ring. Martini⁶ gives several practical examples showing how to tailor groove dimensions to maintain a given squeeze for the operating temperature.

Material Selection/Chemical Compatibility

Seal compounds must work properly over the required temperature range, have the proper hardness to resist extrusion while effectively sealing, and must resist chemical attack and resultant swelling caused by the operating fluids. Table 2 summarizes the most important elastomers, their working temperature range, and their resistance to a range of common working fluids.

Rotary Applications

O-rings are also used to seal rotary shafts where surface speeds and pressures are relatively low. One factor that must be carefully considered when applying O-ring seals to rotary applications is the Gow-Joule effect.⁶ When a rubber O-ring is stretched slightly around a rotating shaft

(e.g., put in tension), friction between the ring and shaft generates heat, causing the ring to contract, exhibiting a negative expansion coefficient. As the ring contracts, friction forces increase, generating additional heat and further contraction. This positive-feedback cycle causes rapid seal failures. Similar failures in reciprocating applications and static applications are unusual because surface speeds are too low to initiate the cycle. Further, in reciprocating applications the seal is moved into contact with cooler adjacent material. To prevent the failure cycle, O-rings are not stretched over shafts but are oversized slightly (circumferentially) and compressed into the sealing groove. The precompression of the cross section results in O-ring stresses that oppose the contraction stress, preventing the failure cycle described. Martini⁶ provides guidelines for specifying the O-ring seal. Following appropriate techniques O-ring seals have run for significant periods of time at speeds up to 750 fpm and pressures up to 200 psi.

2.3 Packings and Braided Rope Seals

Rope packings used to seal stuffing boxes and valves and prevent excessive leakage can be traced back to the early days of the Industrial Revolution. An excellent summary of types of rope seal packings is given in Ref. 7. Novel adaptations of these seal packings have been required as temperatures have continued to rise to meet modern system requirements. New ceramic materials are being investigated to replace asbestos in a variety of gasket and rope-packing constructions.

Materials

Packing materials are selected for the intended temperature and the chemical environment. Graphite-based packing/gaskets are rated for up to 1000°F for oxidizing environments and up to 5400°F for reducing environments.⁸ Used within its recommended temperature, graphite will provide a good seal with acceptable ability to track joint movement during temperature/pressure excursions. Graphite can be laminated with itself to increase thickness or with metal/plastic to improve handling and mechanical strength. Table 2 provides working temperatures for conventional [e.g., nitrile, polytetrafluoroethylene (PTFE), neoprene, among others] gasket/packings. Table 3 provides typical maximum working temperatures for high-temperature gasket/packing materials.

Table 3 Gasket/Rope Seal Materials

Fiber Material	Maximum Working Temperature, °F
<i>Graphite</i>	
Oxidizing environment	1000
Reducing	5400
Fiberglass (glass dependent)	1000
<i>Superalloy metals</i>	
(depending on alloy)	1300–1600
<i>Oxide ceramics^a</i>	
62% Al ₂ O ₃ 24% SiO ₂ 14% B ₂ O ₃ (Nextel 312)	1800 ^b
70% Al ₂ O ₃ 28% SiO ₂ 2% B ₂ O ₃ (Nextel 440)	2000 ^b
73% Al ₂ O ₃ 27% SiO ₂ (Nextel 550)	2100 ^b

^aT. L. Tompkins, "Ceramic Oxide Fibers: Building Blocks for New Applications," Ceramic Industry Publications, Business News Publishing, April 1995.

^bTemperature at which fiber retains 50% (nominal) room temperature strength. Materials can be used at higher temperatures than these for short term. (Consult the manufacturer for guidance.)

Packings and Braided Rope Seals for High-Temperature Service

High-temperature packings and rope seals are required for a variety of applications, including sealing: furnace joints and locations within continuous casting units (gate seals, mold seals, runners, spouts, etc.), among others. High-temperature packings are used for numerous aerospace applications, including turbine casing and turbine engine locations, Space Shuttle thermal protection systems, and nozzle joint seals.

Aircraft engine turbine inlet temperatures and industrial system temperatures continue to climb to meet aggressive cycle thermal efficiency goals. Advanced material systems, including monolithic/composite ceramics, intermetallic alloys (i.e., nickel aluminide), and carbon-carbon composites, are being explored to meet aggressive temperature, durability, and weight requirements. Incorporating these materials in the high-temperature locations in the system, designers must overcome materials issues, such as differences in thermal expansion rates and lack of material ductility.

Designers are finding that one way to avoid cracking and buckling of the high-temperature brittle components rigidly mounted in their support structures is to allow relative motion between the primary and supporting components.⁹ Often this joint occurs in a location where differential pressures exist, requiring high-temperature seals. These seals or packings must exhibit the following important properties: operate hot ($\geq 1300^{\circ}\text{F}$); exhibit low leakage; resist mechanical scrubbing caused by differential thermal growth and acoustic loads; seal complex geometries; retain resilience after cycling; and support structural loads.

In an industrial seal application, a high-temperature all-ceramic seal is being used to seal the interface between a low-expansion-rate primary structure and the surrounding support structure. The seal consists of a dense uniaxial fiber core overbraided with two two-dimensional braided sheath layers.⁹ Both core and sheath are composed of $8\text{-}\mu\text{m}$ alumina-silica fibers (Nextel 550) capable of withstanding $2000+^{\circ}\text{F}$ temperatures. In this application over a heat/cool cycle, the support structure moves 0.3 in. relative to the primary structure,

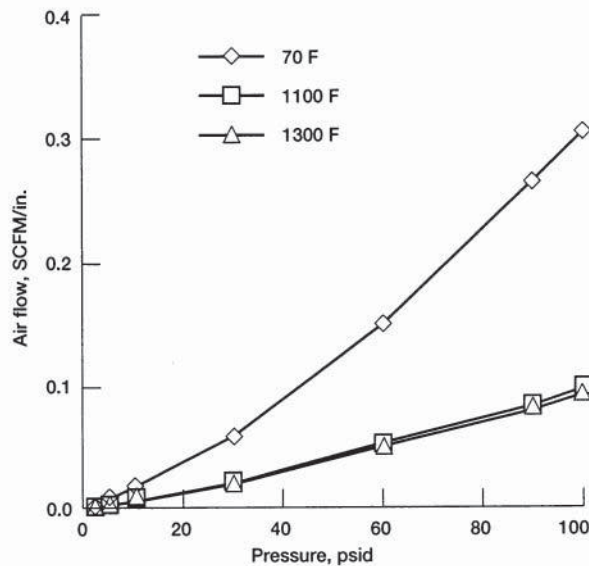


Figure 5 Flow-pressure data for three temperatures, $1/16$ -in.-diameter all-ceramic seal, 0.022 in. seal compression, after scrubbing. (Source: From Ref. 9, "Effects of Compression, Staging and Braid Angle on Braided Rope Seal Performance" by B. M. Steinetz and M. L. Adams; reprinted by permission of the American Institute of Aeronautics and Astronautics, Inc.)

precluding normal fixed-attachment techniques. Leakage flows for the all-ceramic seal are shown in Fig. 5 for three temperatures after simulated scrubbing⁹ (10 cycles \times 0.3 in. at 1300°F). Studies⁹ have shown the benefits of high sheath braid angle and double-stage seals for reducing leakage. Increasing hybrid seal sheath braid angle and increasing core coverage led to increased compressive force (for the same linear seal compression) and one-third the leakage of the conventional hybrid design. Adding a second seal stage reduced seal leakage 30% relative to a single stage.

In a turbine vane application, the conventional braze joint is replaced with a floating-seal arrangement incorporating a small-diameter ($1/16$ -in.) rope seal (Fig. 6). The seal is designed to serve as a seal and a compliant mount, allowing relative thermal growth between the high-temperature turbine vane and the lower temperature support structure, preventing thermal strains and stresses. A hybrid seal consisting of a dense uniaxial ceramic core (8- μ m alumina–silica Nextel 550 fibers) overbraided with a superalloy wire (0.0016-in.-diameter Haynes 188 alloy) abrasion-resistant sheath has proven successful for this application.¹⁰ Leakage flows for the hybrid seal are shown in Fig. 7 for two temperatures and pressures under two preload conditions after simulated scrubbing (10 cycles \times 0.3 in. at 1300°F). Researchers at NASA Glenn Research Center continue to strive for higher operating hybrid seals (metallic sheath over a ceramic fiber core). Recent oxidation studies¹¹ showed that wires made from alumina forming scale base alloys (e.g., Plansee PM2000) could resist oxidation at temperatures to 2200°F (1200°C) for test times up to 70 h. Tests showed that alumina-forming alloys with reactive element additions performed best at 2200°F under all test conditions in the presence of oxygen, moisture, and temperature cycling. Wire samples exhibited slow-growing oxide and adherent scales.

Space Shuttle rocket motor designers have found success in implementing braided carbon ropes as thermal barriers to protect temperature-sensitive downstream O-ring seals.^{12–18} In this application, carbon fiber thermal barriers are used to effectively block the 5500°F gas temperatures from reaching the downstream Viton O-ring pressure seals that are rated

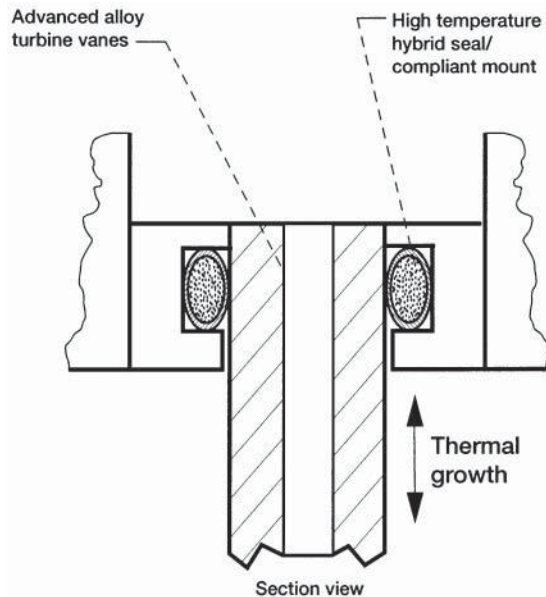


Figure 6 Schematic of turbine vane seal. (Source: From Ref. 10, “High Temperature Braided Rope Seals for Static Sealing Applications” by B. M. Steinetz, et al; reprinted by permission of the American Institute of Aeronautics and Astronautics, Inc.)

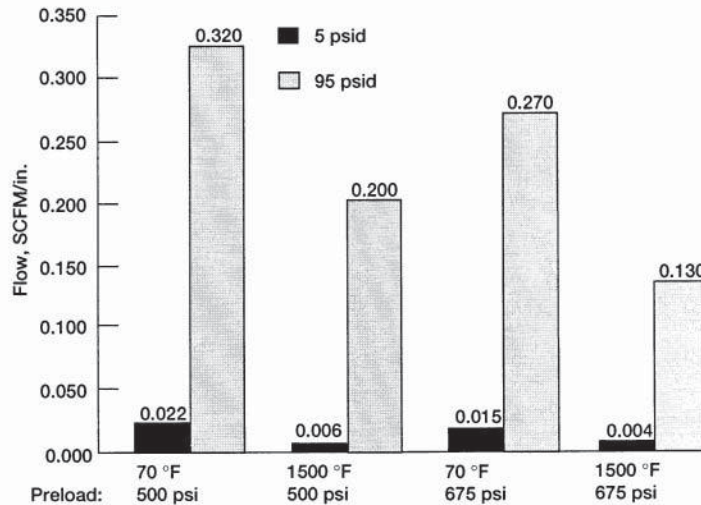


Figure 7 Effect of temperature, pressure, and representative compression on seal flow after cycling for 0.060-in. hybrid vane seal. (Source: From Ref. 10, “High Temperature Braided Rope Seals for Static Sealing Applications” by B. M. Steinetz, et al; reprinted by permission of the American Institute of Aeronautics and Astronautics, Inc.)

to only 500°F (800°F short term). Carbon can be used in this application because of the reducing environment. The braided rope seals are unique in that they effectively block the high temperatures but do not impede the system pressures (900 psi) from seating the O-rings that form the final pressure seal. In the full-scale solid rocket motor tests performed, motor manufacturer ATK-Thiokol measured gas temperatures upstream (hot side) and downstream of the thermal barriers. ATK-Thiokol observed a significant drop in gas temperature across the two thermal barriers, permitting only cool (<200°F) gas to reach the downstream Viton O-rings seals. Full-scale motor tests have certified the carbon thermal barrier for Space Shuttle flight. The NASA Glenn thermal barrier is also being used by other solid rocket manufacturers. Attributes of the thermal barrier include the following:

- *Unique Structure.* The unique braided structure¹⁸ permits designers to tailor the thermal barrier/seal’s properties. Tighter, denser braids form a more effective flow restriction. Looser braids offer more flexibility and allow tighter bend radii.
- *Flexibility/Resiliency.* The carbon thermal barrier provides much-needed flexibility and resiliency to accommodate either joint closings or openings during rocket pressurization and launch not afforded by competing approaches.
- *Self-Seating Feature.* Tests have shown that, upon joint pressurization, the thermal barrier seats itself in the groove to provide a more effective barrier to hot-gas flow.
- *Gas Jet Diffusion.* The thermal barrier diffuses and spreads the incoming high-pressure (900-psi) combustion gas jets, preventing damage to downstream O-rings.
- *Burn Resistance.* NASA Glenn tests showed that the carbon thermal barrier exhibits burn resistance over 60 times greater than similarly constructed ceramic thermal barriers.
- *Slag Block.* The thermal barrier also blocks molten alumina (3700°F) slag (products of combustion) from impinging on temperature-sensitive O-rings.
- *Simplified Installation.* The thermal barrier installs easily into joints in one-sixth the time, eliminating current laborious, time-consuming steps of applying the formerly used joint fill compound, checking joint fill integrity, and replacing/repairing joint fill.

3 DYNAMIC SEALS

3.1 Initial Seal Selection

An engineer approaching a dynamic seal design has a wide range of seals from which to choose. A partial list of seals available ranges from the mechanical face seal through the labyrinth and brush seal, as indicated in Fig. 8. To aid in the initial seal selection, a “decision tree” has been proposed by Fern and Nau.¹⁹ The decision tree (see Fig. 9) has been updated for the current work to account for the emergence of brush seals. In this chart, a majority of answers either “yes” or “no” to the questions at each stage lead the designer to an appropriate seal starting point. If answers are equally divided, both alternatives should be explored using other design criteria, such as performance, size, and cost.

The scope of this chapter does not permit treatment of every entry in the decision tree. However, several examples are given below to aid in understanding its use.

Radial lip seals are used to prevent fluids, normally lubricated, from leaking around shafts and their housings. They are also used to prevent dust, dirt, and foreign contaminants from entering the lubricant chamber. Depending on conditions, lip seals have been designed to operate at very high shaft speeds (6000–12,000 rpm) with light oil mist and no pressure in a clean environment. Lip seals have replaced mechanical face seals in automotive water pumps at pressures to 30 psi, temperatures of -45 – 350°F , and shaft speeds to 8000 sfpm (American Variseal, 1994). Lip seals are also used in completely flooded low-speed applications or in muddy environments. A major advantage of the radial lip seal is its compactness. A 0.32-in. \times 0.32-in. lip seal provides a very good seal for a 2-in.-diameter shaft.

Mechanical face seals are capable of handling much higher pressures and a wider range of fluids. Mechanical face seals are recommended over brush seals where very high pressures must be sealed in a single stage. Mechanical face seals have a lower leakage than brush seals because their effective clearances are several times smaller. However, the mechanical face seal requires much better control of dimensions and tolerates less shaft misalignment and runout, thereby increasing costs.

Turbine Engine Seals. Readers interested particularly in turbine engine seals are referred to Steinetz and Hendricks,²⁰ who review in greater depth the trade-offs in selecting seals for turbine engine applications. Technical factors increasing seal design complexity for aircraft engines include high temperatures ($\geq 1000^{\circ}\text{F}$), high surface speeds (up to 1500 fps), rapid thermal/structural transients, small-space claim, maneuver and landing loads, and the requirement to be lightweight.

3.2 Mechanical Face Seals

The primary elements of a conventional spring-loaded mechanical face seal are the primary seal (the main sealing faces), the secondary seal (seals shaft leakage), and the spring or bellows element that keep the primary seal surfaces in contact, shown in Fig. 8. The primary seal faces are generally lapped to demanding surface flatness, with surface flatness of $40\ \mu\text{in.}$ ($1\ \mu\text{m}$) not uncommon. Surface flatness this low is required to make a good seal, since the running clearances are small. Conventional mechanical face seals operate with clearances of 40 – $200\ \mu\text{in.}$ Dry-running, noncontacting gas face seals that use spiral-groove face geometry reliably run at pressures of 1800 psig and speeds up to 590 fps.^{20a}

Seal Balance

Seal balancing is a technique whereby the primary seal front and rear areas are used to minimize the contact pressure between the mating seal faces to reduce wear and to increase the operating pressure capability. The concept of seal balancing is illustrated in Fig. 10.²¹ The front and rear

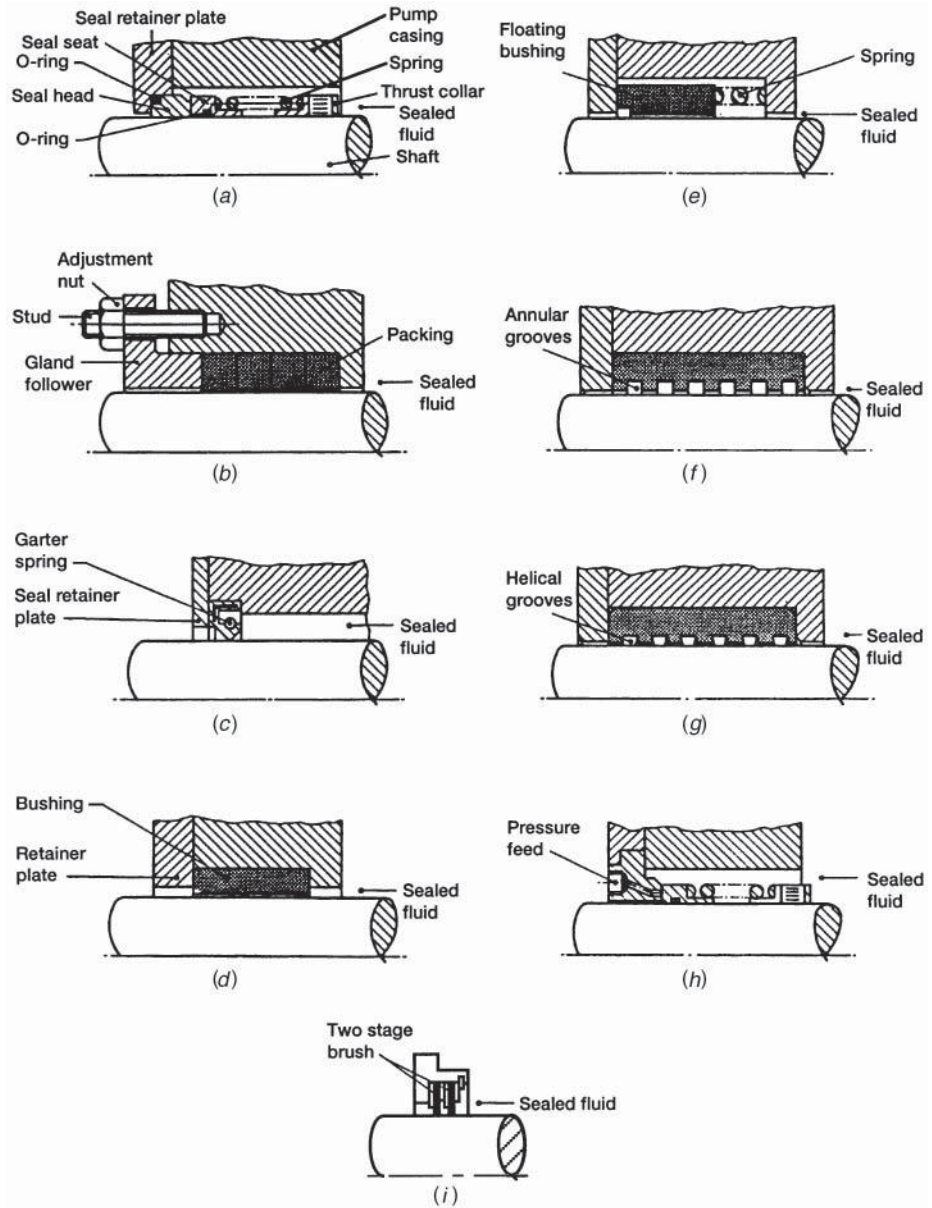


Figure 8 Examples of main types of rotary seal: (a) mechanical face seal; (b) stuffing box; (c) lip seal; (d) fixed bushing; (e) floating bushing; (f) labyrinth; (g) viscoseal; (h) hydrostatic seal; (i) brush seal. (Source: (a)–(h) From Ref. 19, Design Council/University of Brighton Design Archives.)

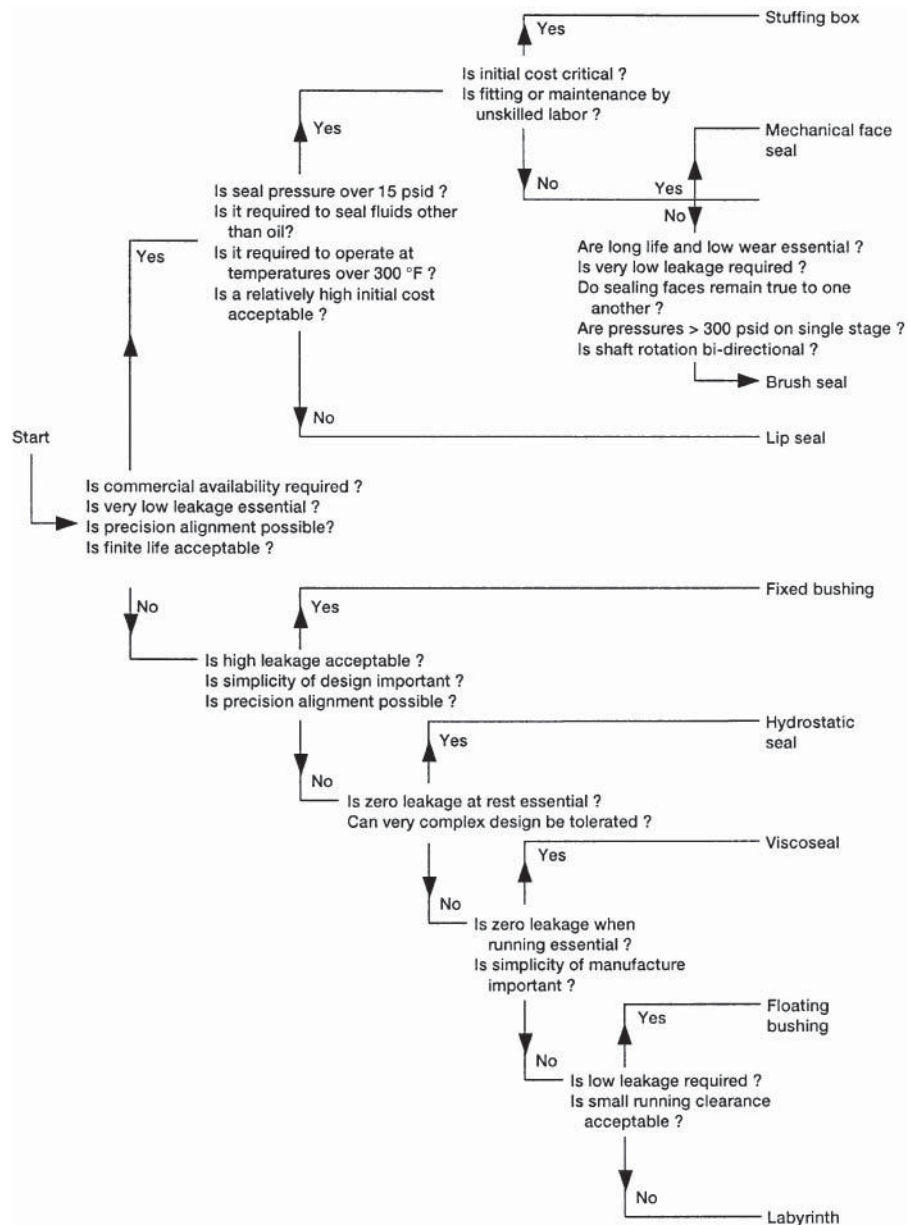


Figure 9 Seal selection chart. (Source: Adapted from Ref. 19, Design Council/University of Brighton Design Archives.)

faces of the seal in Fig. 10a are identical and the full fluid pressure exerted on A' is carried on the seal face A. By modifying the geometry of the primary seal head ring to establish a smaller frontal area A' (Fig. 10b) and to provide a shoulder on the opposite side of the seal ring to form a front face B' , the hydraulic pressure counteracts part of the hydraulic loading from A' . Consequently, the remaining face pressure in the contact interface is significantly reduced. Depending

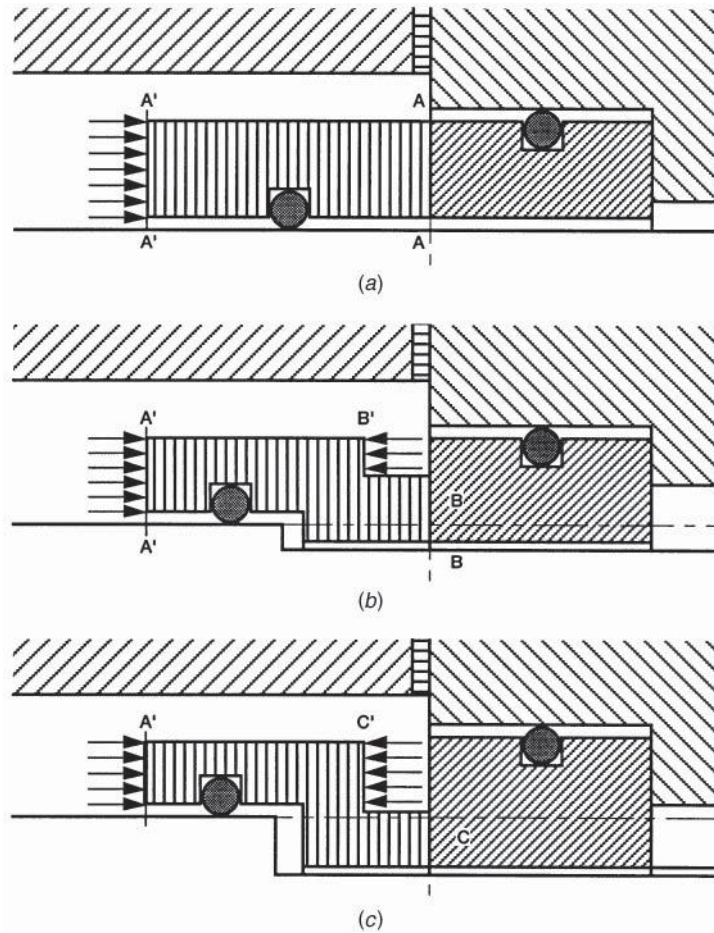


Figure 10 Illustration of face seal balance conditions: (a) unbalanced; (b) partially balanced; (c) fully balanced. (Source: From Ref. 21.)

on the relative sizes of surfaces A' and B' , the seal is either partially balanced (Fig. 10b) or fully balanced (Fig. 10c). In fully balanced seals, there is no net hydraulic load exerted on the seal face. Seals are generally run with a partial balance, however, to minimize face loads and wear while keeping the seal closed during possible transient overpressure conditions. Partially balanced seals can run at pressures greater than six times unbalanced seals can for the same speed and temperature conditions.

Mechanical Face Seal Leakage

Liquid Flow. Minimizing leakage between seal faces is possible only through maintaining small clearances. Volumetric flow (Q) can be determined for the following two conditions²²:

For coned faces:

$$Q = \frac{\pi \phi r_m}{3\mu} \left(\frac{P_o - P_i}{1/h_o^2 - 1/h_i^2} \right)$$

For parallel faces:

$$Q = \frac{-\pi r_m h^3}{6\mu} \frac{(P_o - P_i)}{(r_o - r_i)} \quad h_o = h_i \quad \text{and} \quad (r_o - r_i)/r_m < 0.1$$

where ϕ (radians) is the cone angle (positive if faces are convergent traveling inward radially); r_o , r_i (in.) outer and inner radii; r_m (in.) mean radius (in.); h_o , h_i (in.) outer and inner film thicknesses; P_o , P_i (psi) outer and inner pressures; μ (lb \cdot s/in.²) viscosity. The need for small clearances is demonstrated by noting that doubling the film clearance, h , increases the leakage flow eightfold.

Gas Flow. Closed-form equations for gas flow through parallel faces can be written only for conditions of laminar flow (Reynolds number < 2300). For laminar flow with a parabolic pressure distribution across the seal faces, the mass flow is given as²²

$$\dot{m} = \frac{\pi r_m h^3}{12\mu RT} \frac{(P_i^2 - P_o^2)}{(r_o - r_i)} \quad (r_o - r_i)/r_m < 0.1$$

where R is the gas constant (53.3 lb \cdot ft/lb $_m$ \cdot ^oR for air) and T ($^{\circ}$ R) is the gas temperature (isothermal throughout).

In cases where flow is both laminar and turbulent, iterative schemes must be employed. See Refs. 22 and 23 for numerical algorithms to use in solving for the seal leakage rates. Reference 24 treats the most general case of two-phase flow through the seal faces.

Seal Face Flatness

In addition to lapping faces to the 40- μ in. flatness, there are several other points to consider. The lapped rings should be mounted on surfaces that are themselves flat. The ring must be stiff enough to resist distortions caused either by thermal or fluid pressure stresses.

The primary mode of distortion of a mechanical seal face under combined fluid and thermal stresses is solid-body rotation about the seal's neutral axis.¹⁹ If the sum of the moments M (in.-lb/in.) per unit of circumference around the neutral axis can be calculated, then the angular deflection θ (radians) of the sealing face can be obtained from

$$\theta = \frac{Mr_m^2}{EI}$$

where E (psi) = Young's modulus

I (in.⁴) = second moment of areas about neutral axis

r_m (in.) = mean radius of seal ring

Face Seal Materials

Selecting the correct materials for a given seal application is critical to ensuring desired performance and durability. Seal components for which material selection is important from a tribology standpoint are the stationary nosepiece (or primary seal ring) and the mating ring (or seal seat). Properties considered ideal for the primary seal ring are shown below.²⁵

1. Mechanical

- a. High modulus of elasticity
- b. High tensile strength
- c. Low coefficient of friction
- d. Excellent wear characteristics and hardness
- e. Self-lubrication

2. Thermal
 - a. Low coefficient of expansion
 - b. High thermal conductivity
 - c. Thermal shock resistance
 - d. Thermal stability
3. Chemical
 - a. Corrosion resistance
 - b. Good wettability
4. Miscellaneous
 - a. Dimensional stability
 - b. Good machinability and ease of manufacture
 - c. Low cost and ready availability

Carbon-graphite is often the first choice for one of the running seal surfaces because of its superior dry-running (i.e., startup) behavior. It can run against itself, metals, or ceramics without galling or seizing. Carbon-graphite is generally impregnated with resin or with a metal to increase thermal conductivity and bearing characteristics. In cases where the seal will see considerable abrasives, carbon may wear excessively and then it is desirable to select very hard face seal materials. A preferred combination for very long wear (subject to other constraints) is tungsten carbide running on tungsten carbide. For a comprehensive coverage of face seal material selection, including chemical compatibility, see Ref. 26. Secondary seals are either O-rings or bellows. Temperature ranges and chemical compatibility for common O-ring secondary seals such as nitrile, fluorocarbon (Viton), and PTFE (Teflon) are provided in Table 2.

A mechanical seal is considered to have failed if the seal leakage exceeds the plant site operating or environmental limits. Seal failures can be a major contributor to rotary equipment failure and downtime. The Fluid Sealing Association has compiled an excellent guide for troubleshooting mechanical seals.²⁷ Seals can fail from one or more of the following reasons: (1) incorrect selection of the seal design or materials for the intended application; (2) abuse of the seal before installation; (3) erroneous installation; (4) improper startup, including dry running or failure of the environmental controls; (5) improper equipment operation; (6) contamination of the sealing fluid with either abrasive or corrosive materials; (7) equipment-induced excessive shaft run-out, deflection, vibration, or worn bearings; and (8) a worn-out seal. The effect of bearing performance on seal life is detailed in the *Pump and Systems Handbook*.²⁸

3.3 Emission Concerns

Mechanical face seals have played and will continue to play a major role for many years in minimizing emissions to the atmosphere. New federal, state, and local environmental regulations have intensified the focus on mechanical face seal performance in terms of emissions. Within a short time, regulators have gone from little or no concern about fugitive hazardous emissions to a position of severely restricting all hazardous emissions. For instance, under the authority of Title III of the 1990 Clean Air Act Amendment (CAAA), the U.S. Environmental Protection Agency (EPA) adopted the National Emission Standards for Hazardous Air Pollutants (NESHAP) for the control of emissions of volatile hazardous air pollutants (Ref. STLE, 1994).²⁹ Leak definitions per the regulation [EPA HON Subpart H (5)] are as follows:

Phase I: 10,000 parts per million volumetric (ppmv), beginning on compliance date

Phase II: 5000 ppmv, 1 year after compliance date

Phase III: 1000–5000 ppmv, depending on application, 2 1/2 years after compliance date

The Clean Air Act regulations require U.S. plants to reduce emissions of 189 hazardous air pollutants by 80% in the next several years.³⁰ The American Petroleum Industry (API) has responded with a standard of its own, known as API 682, that seeks to reduce maintenance costs and control volatile organic compound (VOC) emissions on centrifugal and rotary pumps in heavy service. API 682, a pump shaft sealing standard, is designed to help refinery pump operators and similar users to comply with environmental emissions regulations. These regulations will continue to have a major impact on users of valves, pumps, compressors, and other processing devices. Seal users are cautioned to check with their state and local air quality control authorities for specific information.

Sealing Approaches for Emissions Controls

The Society of Tribologists and Lubrication Engineers published a guideline of mechanical seals for meeting the fugitive emissions requirements.²⁹ Seal technology available meets approximately 95% of current and anticipated federal, state, and local emission regulations. Applications not falling within the guidelines include food, pharmaceutical, and monomer-type products where dual seals cannot be used because of product purity requirements and chemical reaction of dual-seal buffer fluids with the sealed product. Bowden³¹ compares various seal arrangements and factors to consider in a design when targeting low fugitive emissions based on operational experience in industrial applications.

Three sealing approaches for meeting the new regulatory requirements are discussed below: single seals, tandem seals, and double seals.²⁹

Single Seals. The most economical approach available is the single seal mounted inside a stuffing box (Fig. 11). Generally, this type of seal uses the pumped product for lubrication. Due to some finite clearance between the faces, there is a small amount of leakage to the atmosphere. Using current technology in the design of a single seal, emissions can be controlled to 500 ppm based on both laboratory and field test data. Emissions to the atmosphere can be eliminated by venting the atmospheric side to a vapor recovery or disposal system. Using this approach, emission readings approaching zero can be achieved. Since single seals have a minimum of contacting parts and normally require minimum support systems, they are considered highly reliable.

Tandem Seals. Tandem seals consist of two seal assemblies between which a barrier fluid operates at a pressure less than the pumped process pressure. The inboard primary seal seals the full

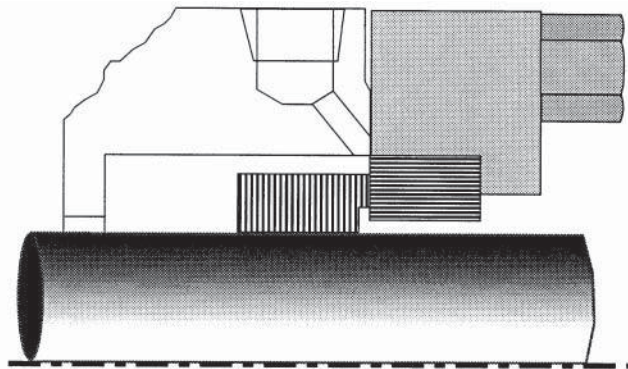


Figure 11 Single seal. (Source: From Ref. 29, With permission from Society of Tribologist and Lubrication Engineers.)

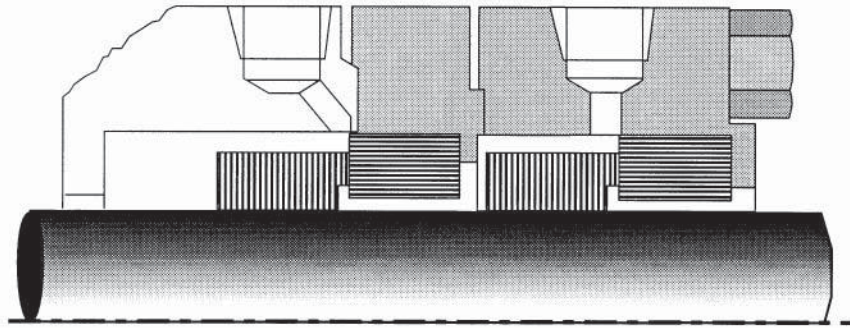


Figure 12 Tandem seal. (Source: From Ref. 29, With permission from Society of Tribologist and Lubrication Engineers.)

pumped product pressure, and the outboard seal typically seals a nonpressurized barrier fluid (Fig. 12). Tandem seal system designs are available that provide zero emission of the pumped product to the environment, provided the vapor pressure of the product is higher than that of the barrier fluid and the product is immiscible in the barrier fluid. The barrier fluid isolates the pumped product from the atmosphere and is maintained by a support system. This supply system generally includes a supply tank assembly and optional cooling system and means for drawing off the volatile component (generally at the top of the supply tank). Examples of common barrier fluids are found in Table 4.

Tandem seal systems also provide a high level of sealing and reliability and are simple systems to maintain, due to the typical use of nonpressurized barrier fluid. Pumped product contamination by the barrier fluid is avoided since the barrier fluid is at a lower pressure than the pumped product.

Double Seals. Double seals differ from tandem seals in that the barrier fluid between the primary and outboard seal is pressurized (Fig. 13). Double seals can be either externally or internally pressurized. An externally pressurized system requires a lubrication unit to pressurize the barrier fluid above the pumped product pressure and to provide cooling. An internally pressurized double seal refers to a system that internally pressurizes the fluid film at the inboard faces as

Table 4 Properties of Common Barrier Fluids for Tandem or Double Seals^a

Barrier Fluid	Temperature Limits, °F		Comments
	Lower	Upper	
Water	40	180	Use corrosion-resistant materials Protect from freezing
Propylene glycol	-76	368	Consult seal manufacturer for proper mixture with water to avoid excessive viscosity
<i>n</i> -Propyl alcohol	-147	157	
Kerosene	0	300	
No. 2 diesel fuel	10	300	Contains additives

^aSTLE, "Guidelines for Meeting Emission Regulations for Rotating Machinery with Mechanical Seals," Special Publication SP-30, Society of Tribologists and Lubrication Engineers, Park Ridge, IL, 1990.

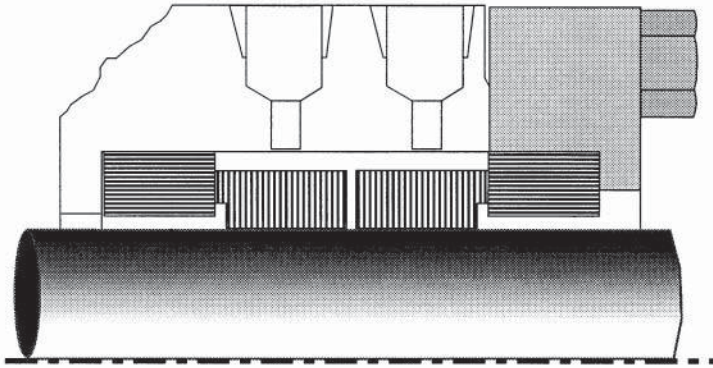


Figure 13 Double seal. (Source: From Ref. 29, With permission from Society of Tribologist and Lubrication Engineers.)

the shaft rotates. In this case, the barrier fluid in the seal chamber is normally at atmospheric pressure. This results in less heat generation from the system.

Application Guide. The areas of application based on emissions to the atmosphere for the three types of seals discussed are illustrated in Fig. 14. The scope of this chart is for seals less than 6 in. in diameter, for pressures 600 psig and less, and for surface speeds up to 5600 fpm. Waterbury³⁰ provides a modern overview of several commercial products aimed at achieving zero leakage or leak-free operation in compliance with current regulations.

3.4 Noncontacting Seals for High-Speed/Aerospace Applications

For very high speed turbomachinery, including gas turbines, seal runner speeds may reach speeds greater than 1300 fps, requiring novel seal arrangements to overcome wear and pressure limitations of conventional face seals. Two classes of seals are used that rely on a thin film of air to separate the seal faces. Hydrostatic face seals port high-pressure fluid to the sealing face to induce opening force and maintain controlled face separation (see Fig. 15). The fluid pressure developed between the faces is dependent upon the gap dimension and the pressure varies between the lower and upper limits shown in the figure. Any change in the design clearance results in an increase or decrease of the opening force in a stabilizing sense. Of the four configurations shown, the coned seal configuration is the most popular. Converging faces are used to provide seal stability. Hydrostatic face seals suffer from contact during startup. To overcome this, the seals can be externally pressurized, but this adds cost and complexity.

The aspirating hydrostatic face seal (Fig. 15d) under development by GE and Stein Seal for turbine engine applications provides a unique failsafe feature.^{32–34} The seal is designed to be open during initial rotation and after system shutdown—the two periods during which potentially damaging rubs are most common. Upon system pressurization, the aspirating teeth set up an initial pressure drop across the seal (6 psi nominal) that generates a closing force to overcome the retraction spring force F_s , causing the seal to close to its operating clearance (nominal 0.0015–0.0025 in.). System pressure is ported to the face seal to prevent touchdown and provide good film stiffness during operation. At engine shutdown, the pressure across the seal drops and the springs retract the seal away from the rotor, preventing contact.

Hydrodynamic or self-acting face seals incorporate lift pockets to generate a hydrodynamic film between the two faces to prevent seal contact. A number of lift pocket configurations are employed, including shrouded Rayleigh step, spiral groove, circular groove, and annular groove (Fig. 16). In these designs, hydrodynamic lift is independent of the seal pressure; it is

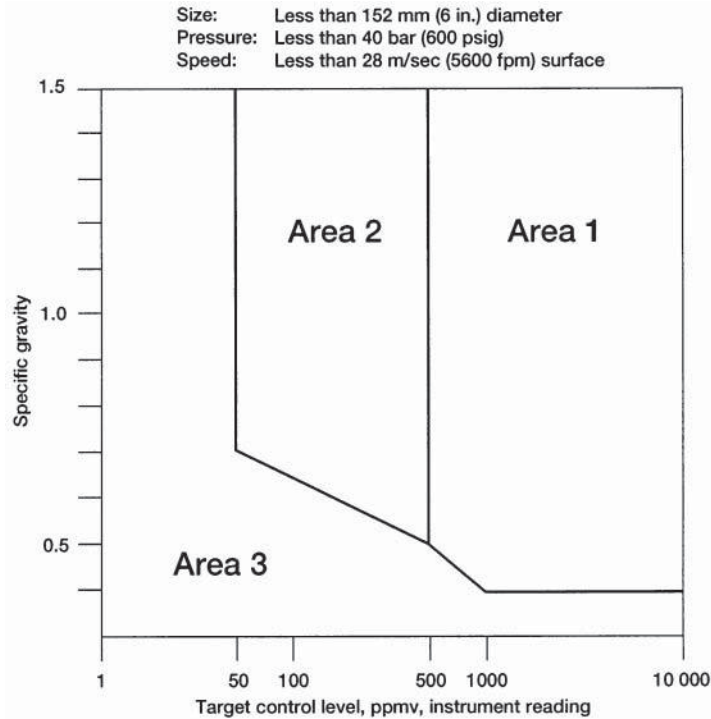


Chart area	Recommended technology
1	General purpose single seals, or dual (double and tandem) seals
2	Special purpose single seals, or dual (double and tandem) seals
3	Dual pressurized (double) seals Single or dual non-pressurized (tandem) seals vented to a closed vent system, above 0.4 specific gravity

Figure 14 Application guide to control emissions. (Source: From Ref. 29, With permission from Society of Tribologist and Lubrication Engineers.)

proportional to the rotation speed and to the fluid viscosity. Therefore a minimum speed is required to develop sufficient lift force for face separation. Hydrodynamic seals operate on small (≤ 0.0005 in. nominal) clearances, resulting in very low leakage compared to labyrinth or brush seals, as shown in Fig. 17.³⁵ Because rubbing occurs during startup and shutdown, seal face materials must be selected for good rubbing characteristics for low wear (see Face Seal Materials, above).

Computer Analysis Tools: Face/Annular Seals

To aid aerospace and industrial seal designers alike, NASA sponsored the development of computer codes to predict the seal performance under a variety of conditions.³⁶ NASA seal design computer codes are available to U.S. persons through Open Channel Software.³⁷ Codes were developed to treat both incompressible (e.g., liquid) and compressible (e.g., gas) flow conditions. In general, the codes assess seal performance characteristics, including load capacity,

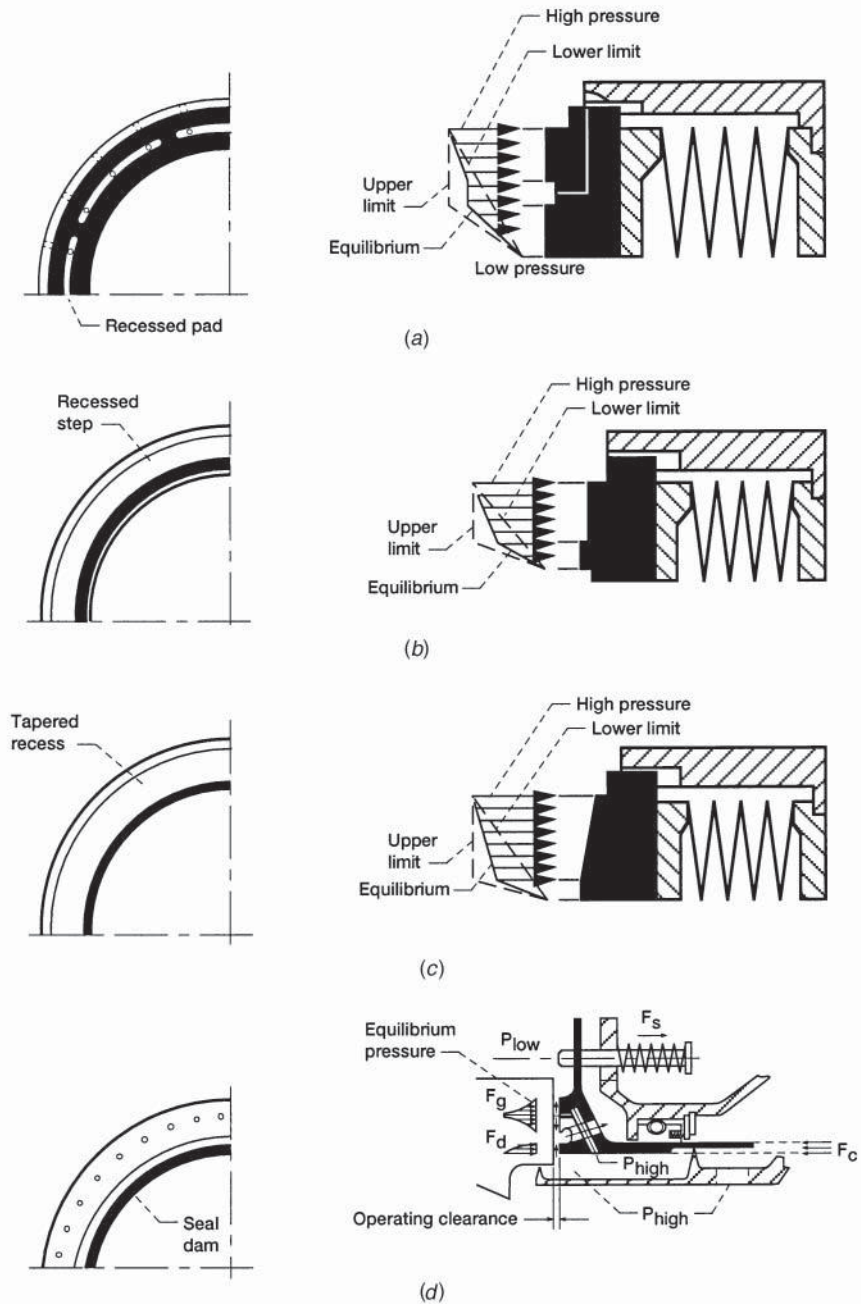


Figure 15 Self-energized hydrostatic noncontacting mechanical face seals: (a) recessed pads with orifice compensation; (b) recessed step; (c) convergent tapered face; (d) aspirating seal. (Source: (a)–(c) from Ref. 1; (d) from Ref. 32, “Testing of a High Performance Compressor Discharge Seal” by J. Munson; reprinted by permission of the American Institute of Aeronautics and Astronautics, Inc.)

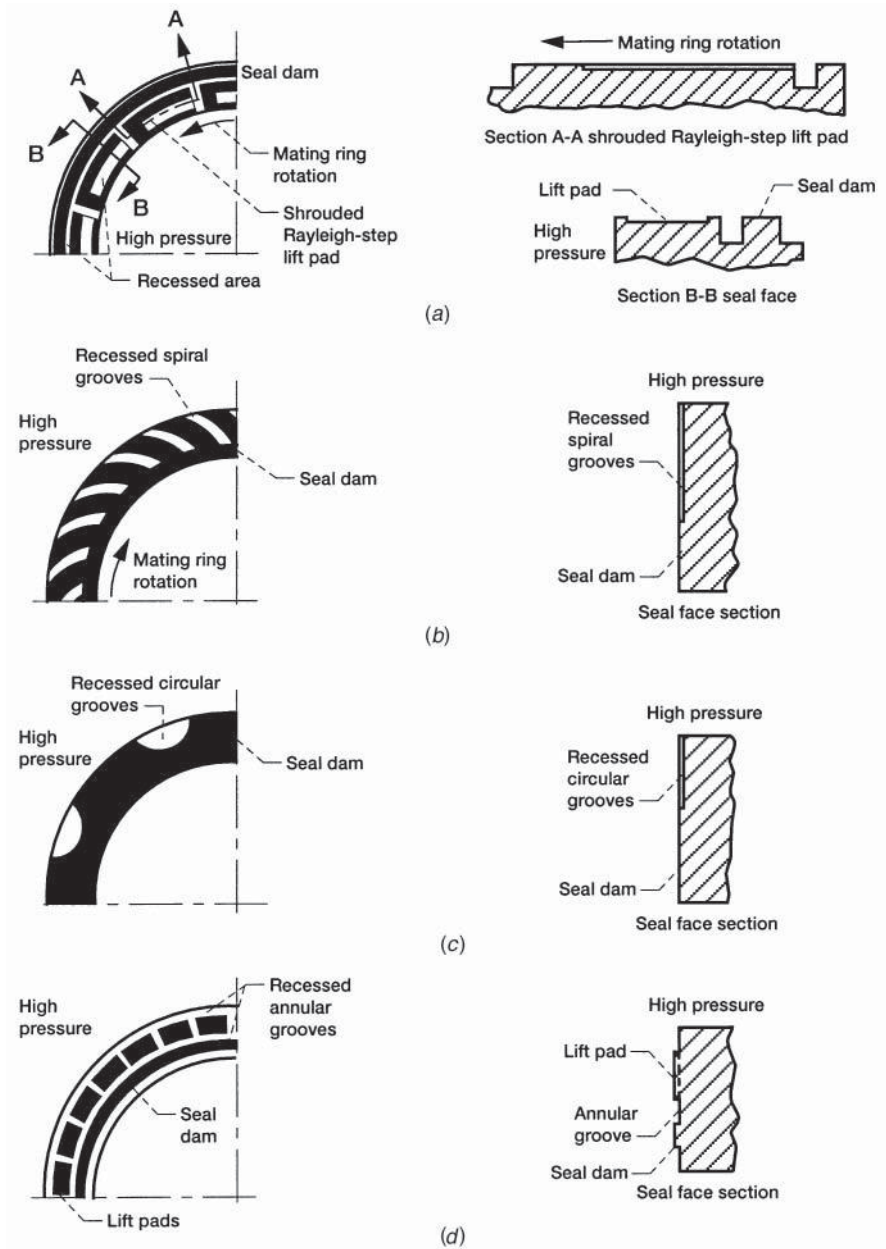


Figure 16 Various types of hydrodynamic noncontacting mechanical face seals: (a) shrouded Rayleigh step; (b) spiral groove; (c) circular groove; (d) annular groove. (Source: From Ref. 1.)

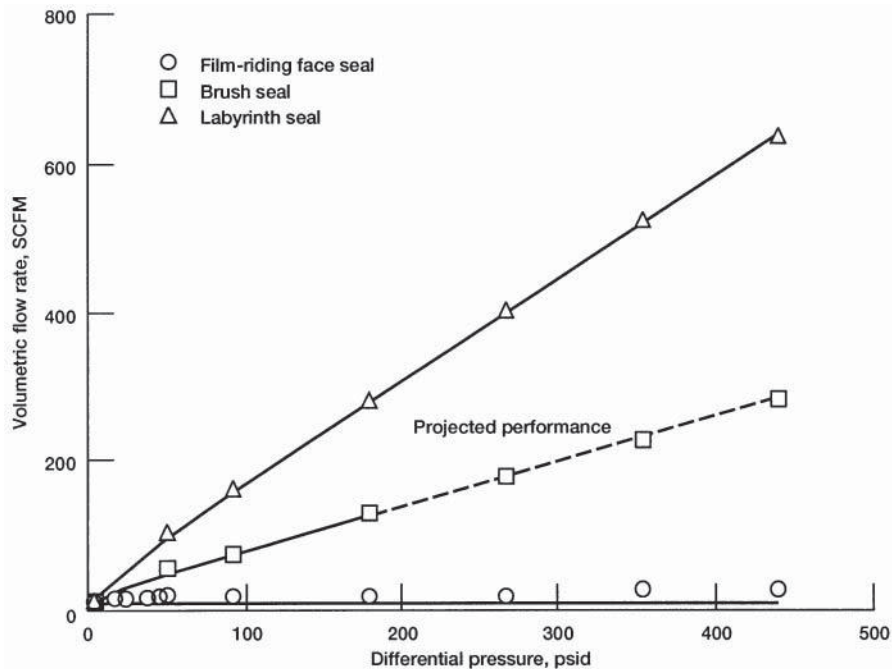


Figure 17 Comparison of brush, labyrinth, and self-acting, film-riding face seal leakage rates as function of differential pressure. Seal diameter, 5.84 in. (Source: From Ref. 35, “Testing of a High Performance Compressor Discharge Seal” by J. Munson; reprinted by permission of the American Institute of Aeronautics and Astronautics, Inc.)

leakage flow, power requirements, and dynamic characteristics in the form of stiffness and damping coefficients. These performance characteristics are computed as functions of seal and groove geometry, loads or film thicknesses, running speed, fluid viscosity, and boundary pressures. The GFACE code predicts performance for the following face seal geometries: hydrostatic, hydrostatic recess, radial and circumferential Rayleigh step, and radial and circumferential tapered land. The GCYLT code predicts performance for both hydrodynamic and hydrostatic cylindrical seals, including the following geometries: circumferential multilobe and Rayleigh step, Rayleigh step in direction of flow, tapered and self-energized hydrostatic. A description of these codes and their validation are given by Shapiro.³⁸ The SPIRALG/SPIRALI codes predict characteristics of gas-lubricated (SPIRALG) and liquid-lubricated (SPIRALI) spiral groove, cylindrical, and face seals.³⁹

Dynamic response of seal rings to rotor motions is an important consideration in seal design. For contact seals, dynamic motion can impose significant interfacial forces, resulting in high wear and reduction in useful life. For fluid-film seals, the rotor excursions are generally greater than the film thickness, and if the ring does not track, contact and failure may occur. The computer code DYSEAL predicts the tracking capability of fluid-film seals and can be used for parametric geometric variations to find acceptable configurations.⁴⁰

3.5 Labyrinth Seals

By their nature, labyrinth seals are clearance seals that also permit shaft excursions without potentially catastrophic rub-induced rotor instability problems. By design, labyrinth seals restrict leakage by dissipating the kinetic energy of fluid flow through a series of flow

constrictions and cavities that sequentially accelerate and decelerate the fluid flow or change its direction abruptly to create the maximum flow friction and turbulence. The ideal labyrinth seal would transform all kinetic energy at each throttling into internal energy (heat) in each cavity. However, in practical labyrinth seals, a considerable amount of kinetic energy is transferred from one passage to the next. The advantage of labyrinth seals is that the speed and pressure capability is limited only by the structural design. One disadvantage, however, is a relatively high leakage rate. Labyrinth seals are used in so many gas-sealing applications because of their very high running speed (1500 ft/s), pressure (250 psi), and temperature ($\geq 1300^\circ\text{F}$) and the need to accommodate shaft excursions caused by transient loads. Labyrinth seal leakage rates have been reduced over the years through novel design concepts but are still higher than desired because labyrinth seal leakage is clearance dependent and this clearance opens due to periodic transient rubs.

Seal Configurations

Labyrinth seals can be configured in many ways (Fig. 18). The labyrinth seal configurations typically used are straight, angled-teeth straight, stepped, staggered, and abradable or wear-in. Optimizing labyrinth seal geometry depends on the given application and greatly affects the labyrinth seal leakage. Stepped labyrinth seals have been used extensively as turbine inter-stage air seals. Leakage flow through inclined, stepped labyrinths is about 40% that of straight labyrinths for similar conditions (Fig. 19). Performance benefits of stepped labyrinths must be balanced with other design issues. They require more radial space, are more difficult to manufacture, and may produce an undesirable thrust load because of the stepped area.

Leakage Flow Modeling

Leakage flow through labyrinth seals is generally modeled as a sequential series of throttlings through the narrow blade tip clearances. Ideally, the kinetic energy increase across each annular orifice would be completely dissipated in the cavity. However, dissipation is not complete. Various authors handle this in different ways: Egli⁴³ introduced the concept of “carryover” to account for the incomplete dissipation of kinetic energy in straight labyrinth seals. Vermes⁴⁴ introduced the residual energy factor, α , to account for the residual energy in the flow as it passes from one stage to the next:

$$W = 5.76K \frac{A_g}{[RT_o]^{1/2}} \frac{P_o}{[1 - \alpha]^{1/2}} \beta \quad \text{where} \quad \beta = \left[\frac{1 - \left[\frac{P_N}{P_o} \right]^2}{N - \ln \left[\frac{P_N}{P_o} \right]} \right]^{1/2}$$

and the residual energy factor

$$\alpha = \frac{8.52}{\left[\frac{TP - L}{c} \right] + 7.23}$$

- where
- A_g = flow area of single annular orifice (in.²)
 - c = clearance (in.)
 - g_c = gravitational constant (32.2 ft/s²)
 - G = mass flux (lbm/ft²·s)
 - K = clearance factor for annular orifice (see Fig. 20)
 - L = tooth width at sealing point (in.)
 - N = number of teeth (in.)
 - N_{Re} = Reynolds number, defined as $G(c/12)/\mu g_c$
 - TP = tooth pitch (in.)
 - P_o, P_N = inlet pressure, pressure at tooth N

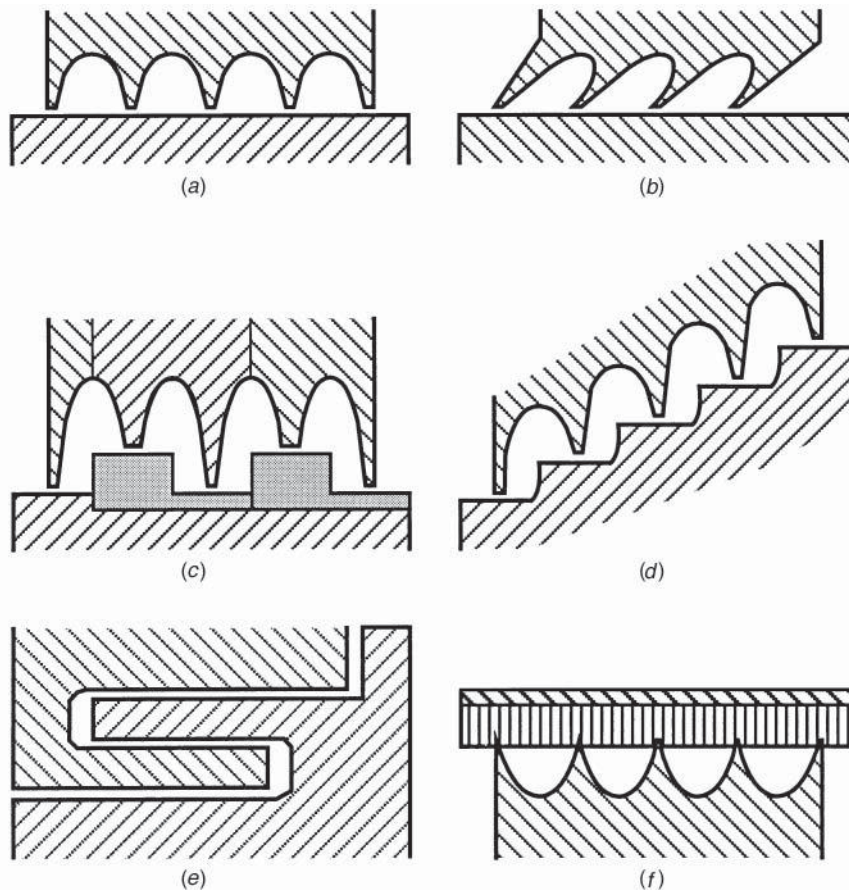


Figure 18 Labyrinth seal configurations: (a) straight labyrinth; (b) inclined- or angled-teeth straight labyrinth; (c) staggered labyrinth; (d) stepped labyrinth; (e) interlocking labyrinth; (f) abrasion-resistant (wear-in) labyrinth. (Source: From Ref. 41.)

R = gas constant (lbf-ft/lbm-°R)

T_o = gas inlet temperature (°R)

W = weight flow lb/s

μ = gas viscosity (lbf-s/ft²)

The clearance factor is plotted in Fig. 20 for a range of Reynolds numbers and tooth width-to-clearance ratios. Since K is a function of N_{Re} and since N_{Re} is a function of the unknown mass flow, the necessary first approximation can be made with $K = 0.67$. Vermes⁴⁴ also presented methods for calculating mass flow for stepped labyrinth seals and for off-design conditions (e.g., the stepped seal teeth are offset from their natural lands). Tooth shape also plays a role in leakage resistance. Mahler⁴⁵ showed that sharp corners provide the highest leakage resistance.

Applications

There are innumerable applications of labyrinth seals in the field. They are used to seal rolling-element bearings, machine spindles, and other applications where some leakage can

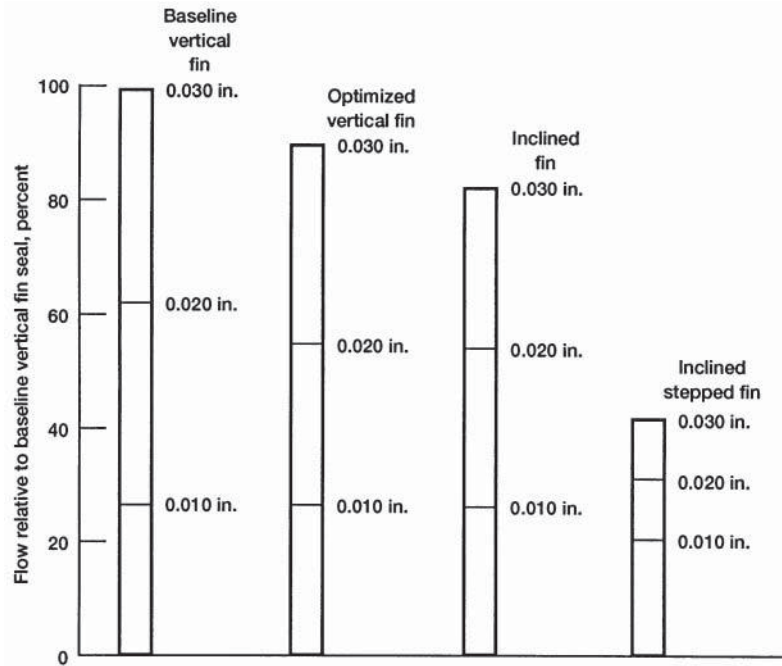


Figure 19 Labyrinth seal leakage flow performance for typical designs and clearances relative to a baseline five-finned straight labyrinth seal of various gaps at pressure ratio of 2. (Source: From Ref. 42. With permission from ASME.)

be tolerated. Since the development of the gas turbine engine, the labyrinth seal has been perhaps the most common seal applied to sealing both primary and secondary air flow.²⁰ Its combined pressure–speed–life limits have for many years exceeded those of its rubbing-contact seal competitors. Labyrinth seals are also used extensively in cryogenic rocket turbopump applications.

Seals and Rotordynamic Stability

Although the primary function of a seal is to control leakage, a secondary but equally important purpose is not to negatively affect rotordynamic stability, especially in high-speed turbomachinery. When a clearance gap such as in either annular or labyrinth seals changes with time, lateral forces occur that act out of phase with case distortion. Depending on changes in gap in the flow direction, excitation or damping of the lateral forces occurs. If the lateral forces become too large, they can contribute to shaft instability problems. These problems and several solutions, including swirl brakes, are further discussed in Hendricks,⁴⁶ Bently,⁴⁷ Alford,^{48–50} and Benckert and Wachter,⁵¹ among others.

Computer Analysis Tools: Labyrinth Seals

The computer code KTK calculates the leakage and pressure distribution through a labyrinth seal based on a detailed knife-to-knife (KTK) analysis. This code was developed by Allison Gas Turbines for the Air Force⁵² and is also documented in Shapiro et al.⁴⁰ Rhode and Nail⁵³ present recent work in the application of a Reynolds-averaged computer code to generic labyrinth seals operating in the compressible region Mach number ≥ 0.3 .

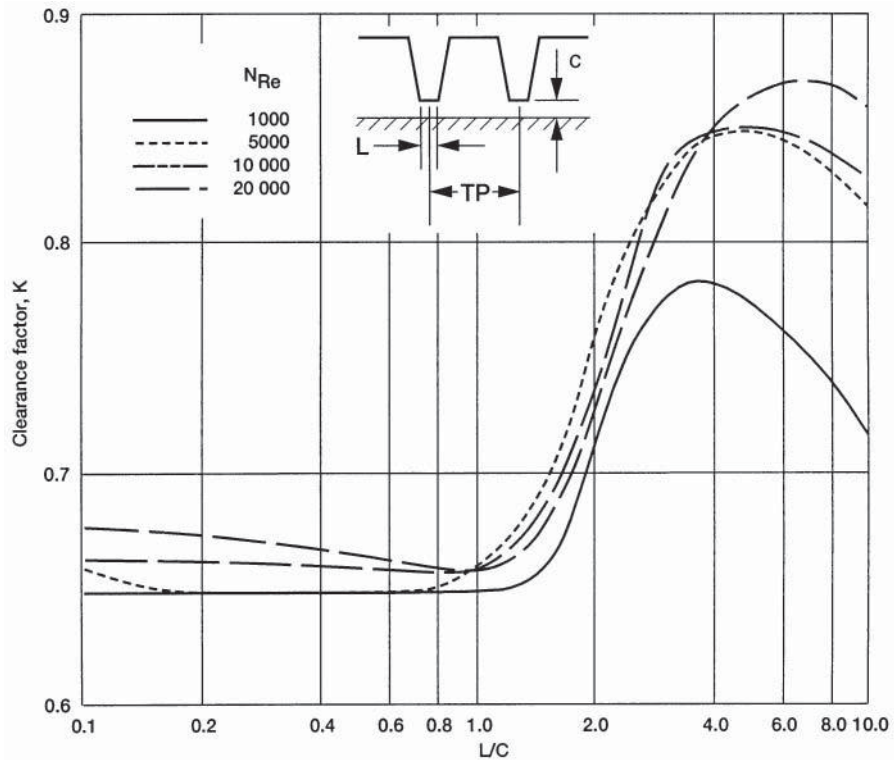


Figure 20 Clearance factor (K) versus ratio of tooth width (L) to tooth clearance (c). [For various Reynolds numbers (N_{Re}).] (Source: From Ref. 44. With permission from ASME.)

3.6 Honeycomb Seals

Honeycomb seals are used extensively in mating contact with labyrinth knife edges machined onto the rotor in applications where there are significant shaft movements. After brazing the honeycomb material to the case, the inner diameter is machined to seal tolerance requirements. Properly designed honeycomb seals, in extensive tests performed by Stocker et al.⁵⁴ under a NASA contract, showed dramatic leakage reductions under select gap and honeycomb cell size combinations.

For applications where low leakage is paramount, designers will specify a small radial clearance between the labyrinth teeth and abradable surface (honeycomb or sprayed abradable). Designers will take advantage of normal centrifugal growth of the rotor to reduce this clearance to line-to-line and often to a wear-in condition, making an effective labyrinth seal. A “green” slow-speed-ramp wear-in cycle is recommended.

Materials. Honeycomb elements are often fabricated of Hastelloy X,⁵⁵ a nickel-based alloy. Honeycomb seals provide for low-energy rubs when transient conditions cause the labyrinth knife edges to wear into the surface (low-energy rubs minimize potentially damaging shaft vibrations). In very high surface speed applications and where temperatures are high the labyrinth teeth are “tipped” with a hard abrasive coating, increasing cutting effectiveness and reducing the thermal stresses in the labyrinth teeth during rubs.

Honeycomb Annular Seals. Honeycomb seals are also being considered now as annular seals to greatly improve damping over either smooth surfaces or labyrinth seals. Childs et al.⁵⁶ showed that honeycombs properly applied in annular seals control leakage, have good stiffness, and exhibit damping characteristics six times those of labyrinth seals alone.

3.7 Brush Seals

As described by Ferguson,⁴² the brush seal is the first simple, practical alternative to the finned labyrinth seal that offers extensive performance improvements. Benefits of brush seals over labyrinth seals include the following:

- Reduced leakage compared to labyrinth seals. If properly applied, leakage reductions upward of 50% are possible.
- Flexible brush seal accommodates shaft excursions due to stop/start operations and other transient conditions. Labyrinth seals often incur permanent clearance increases under such conditions, degrading seal and machine performance.
- Requires significantly less axial space than labyrinth seal.
- More stable leakage characteristics over long operating periods.

Brush seals have matured significantly over the past 20 years. Typical operating conditions of state-of-the-art brush seals include the following⁵⁷:

Differential pressure	Up to 300 psid per stage; higher pressures (e.g., 1000 psid) possible with multiple stages
Surface speed	Up to 1200 ft/s
Operating temperature	Up to 1200°F
Diameter range	Up to 120 in.

Basic brush seal construction is quite simple, as shown in cross section in Fig. 21. A dense pack of fine-diameter wire bristles is sandwiched and welded between a backing ring (downstream side) and a side plate (upstream side). The wire bristles protrude radially inward and are machined to form a brush bore fit around a mating rotor, with a slight interference. Brush seal interferences and preload must be properly selected to prevent potentially catastrophic overheating of the rotor and excessive rotor thermal growths. The weld on the seal outer diameter is machined to form a close-tolerance outer diameter sealing surface that is fitted into a suitable seal housing.

To accommodate anticipated radial shaft movements, the bristles must bend. To allow the bristles to bend without buckling, the wires are oriented at an angle (typically 45°–55°) to a radial line through the shaft. The bristles point in the direction of rotation. The angled construction also greatly facilitates seal installation, considering the slight inner diameter interference with the rotors. The backing ring provides structural support to the otherwise flexible bristles and assists the seal in limiting leakage. To minimize brush seal hysteresis caused by brush bristle binding on the back plate, new features have been added to the backing ring. These include reliefs of various forms. An example design is shown in Fig. 21 and includes the recessed pocket and seal dam. The recessed pocket assists with pressure balancing of the seal and the relatively small contact area at the seal dam minimizes friction, allowing the bristles to follow the speed-dependent shaft growths. Bristle-free radial length and packing pattern are selected to accommodate anticipated shaft radial movements while operating within the wire's elastic range at temperature. A number of brush seal manufacturers^{57,58} include some form of flow deflector (e.g., see flexi-front plate in Fig. 21) on the high-pressure side of the wire bristles. This element aids in mitigating the radial pressure closing loads (e.g., sometimes known as “pressure closing”) caused by air forces urging the bristles against the shaft. This element can also aid in

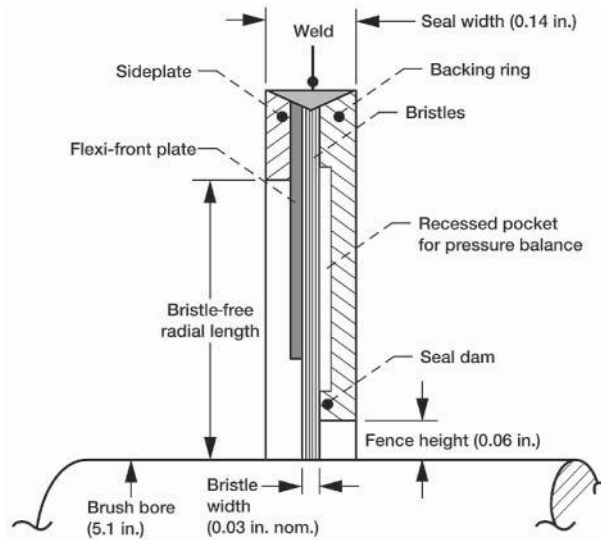


Figure 21 Brush seal cross section with typical dimensions. (Source: From Ref. 20.)

reducing installation damage, bristle flutter in highly turbulent flow fields, and foreign-object damage. The backing ring clearance is sized slightly larger than anticipated rotor radial excursions and relative thermal and mechanical growth to ensure that the rotor never contacts the ring, causing rotor and casing damage. An abrasible rub surface added to the backing ring has been proposed to mitigate this problem by allowing tighter backing-plate clearances.

Brush Seal Design Considerations

To properly design and specify brush seals for an application, many design factors must be considered and traded off. A comprehensive brush seal design algorithm was proposed by Holle and Krishan.⁵⁹ An iterative process must be followed to satisfy seal basic geometry, stress, thermal (especially during transient rub conditions), leakage, and life constraints to arrive at an acceptable design. Table 5 illustrates many of the characteristics that must be considered and understood for a successful brush seal design.⁶⁰ Design criteria are required for each of

Table 5 Brush Seal Characteristics Evaluated during Design for Successful Application

Pressure capability	Seal upstream protection
Frequency	Seal high- and low-cycle fatigue (HCF, LCF) analysis
Seal leakage	Seal oxidation
Seal stiffness	Seal creep
Seal blow-down (e.g., pressure-closing effect)	Seal wear
Bristle tip forces and pressure-stiffening effect	Solid-particle erosion
Seal heat generation	Reverse rotation
Bristle tip temperature	Seal life/long-term considerations
Rotor dynamics	Performance predictions
Rotor thermal stability	Oil sealing
Secondary flow and cavity flow (including swirl flow)	Shaft considerations (e.g., coating)

Source: From Ref. 60.

the different potential failure modes, including stress, fatigue life, creep life, wear life, and oxidation life, among others.

Implementation Issues. Improper design and implementation of brush seals can result in premature seal failures. The following is a partial list of potential pitfalls to be mindful of in specifying brush seals⁶⁰:

- *Excessive Interference.* Specifying too tight of a fit at assembly can lead to excessive frictional heat and wear of bristles. In the worst case, seals can become thermally unstable: Frictional heating can cause the rotor to grow radially into the seal, thereby increasing frictional heating and leading to additional rotor growth. If left unchecked, the rotor can grow into the backing plate, leading to seal and possibly equipment failure. Designers need to consider the relative seal-to-shaft closure across the entire operating speed/temperature range. In large ground-based turbine designs, brush seals are often assembled with a clearance to preclude excessive interference and heating during thermal and speed transients.
- *Excessive Operating Speed.* Brush seals have been run successfully to 1200 fps.^{57,58} In some limited applications they have run faster than this. However, excessive surface speeds combined with high unit pressures can lead to excessive heat generation and premature failure.
- *Inadequate Understanding of Flow Fields.* The design of upstream cavities is key to reducing brush seal flutter, especially in high-swirl-flow fields.⁶⁰ Poor designs lead to bristle aerodynamic instability and high-cycle fatigue. Furthermore, brush seals can restrict leakage so much better than labyrinth seals that inadequate flow may be supplied to expensive downstream components (e.g., turbine vanes, blades, buckets, and/or wheel cavities), resulting in life-limiting conditions for those components. A clear understanding of the flow fields is essential for successful implementation.
- *Improper Brush Pack Design.* Designers should consult with brush seal manufacturers to aid in specifying the brush parameters. Manufacturers can aid in selecting the correct wire diameter, brush pack width, bristle free height, and fence height for the speed, pressure, and transient conditions anticipated. Improper brush pack width, for instance, can result in excessive bending of the bristles under the backing ring, leading to excessive bristle wear.

Brush Pack Considerations. Depending on required sealing pressure differentials and life, wire bristle diameters are chosen in the range of 0.0028–0.006 in.⁶¹ Better load and wear properties are found with larger bristle diameters. Bristle pack widths also vary depending on application: The higher the pressure differential, the greater the pack width. Higher pressure applications require bristle packs with higher axial stiffness to prevent the bristles from blowing under the backing ring. Dinc et al.⁶⁰ have developed brush seals that have operated at air pressures up to 400 psid in a single stage. Brush seals have been made in very large diameters. Large brush seals, especially for ground power applications, are often made segmented to allow easy assembly and disassembly, especially on machines where the shaft stays in place during refurbishment.

Other Considerations. If not properly considered, brush seals can exhibit three other phenomena deserving some discussion: *seal hysteresis*, *bristle stiffening*, and *pressure closing*. As described in Short et al.⁶¹ and Basu et al.,⁶² after the rotor moves into the bristle pack (due to radial excursions or thermal growths), the displaced bristles do not immediately recover

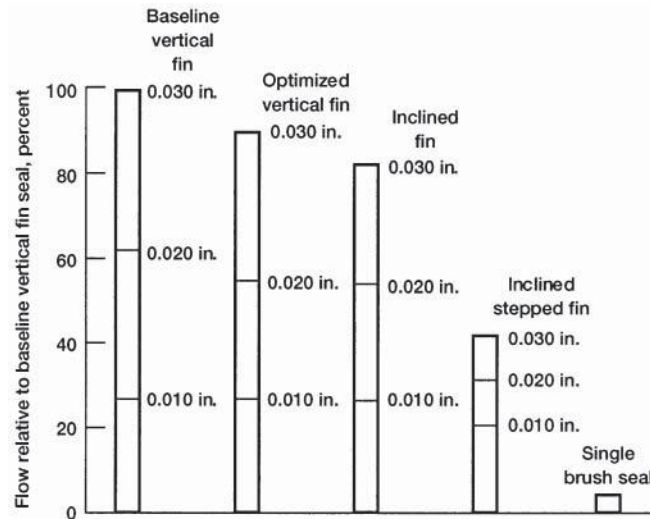


Figure 22 Sealing performance of new brush seals relative to baseline five-finned labyrinth seals of various radial gaps at pressure ratio of 2. (Source: From Ref. 42. With permission from ASME.)

against the frictional forces between them and the backing ring. As a result, a significant leakage increase (more than double) was observed⁶² following rotor movement. This leakage hysteresis exists until after the pressure load is removed (e.g., after the engine is shut down). Furthermore, if the bristle pack is not properly designed, the seal can exhibit a considerable stiffening effect with application of pressure. This phenomenon results from interbristle friction loads, making it more difficult for the brush bristles to flex during shaft excursions. Air leaking through the seal also exerts a radially inward force on the bristles, resulting in what has been termed pressure closing or bristle “blow-down.” This extra contact load, especially on the upstream side of the brush, affects the life of the seal (upstream bristles are worn in either a scalloped or coned configuration) and higher interface contact pressure. Because of these and other considerations, designers should consult with brush seal manufacturers^{57,58} for application assistance.

Multiple brush seals are generally used where large pressure drops must be accommodated. The primary reason for using multiple seals is not to improve sealing but to reduce pressure-induced distortions in the brush pack, namely axial brush distortions under the backing ring, that cause wear. Researchers have noticed greater wear on the downstream brush if the flow jet coming from the upstream brush is not deflected away from the downstream brush-rotor contact.

Leakage Performance Comparisons

Ferguson⁴² compared brush seal leakage with that of traditional five-finned labyrinth seals of various configurations. The results of this study (Fig. 22) indicate that the flow of a *new* brush seal is only 4% that of a vertical finned seal with a 0.03-in. radial gap and one-fifth that of an inclined-fin labyrinth seal with a step up and a 0.01-in. gap.

Addy et al.⁶³ showed similar large reductions in leakage testing a 5.1-in. bore seal across a wide temperature and speed range. Table 6 compares air leakage between a new brush seal and similarly sized labyrinth seals.

Table 6 Comparison of New Labyrinth Seal (Smooth and Honeycomb Lands) versus New Brush Seal Leakage Rates for Comparable Conditions

Seal	Rotor Diameter (in.)	Seal Clearance (in.)	Brush Seal Interference (in.)	Pressure Ratio, P_i / P_o	ϕ -Flow Parameter, $\frac{(\text{lb}_m \cdot \text{R}^{1/2})}{\text{lb}_f \cdot \text{s}}$	Mass Flow, (lb_m / s)
4-Tooth labyrinth vs. smooth land ^a	6.0	0.010	—	3.0	0.36	0.141
4-Tooth labyrinth vs. honeycomb land, 0.062-in. cell size ^a	6.0	0.010	—	3.0	0.35	0.137
Brush seal ^b	5.1	—	0.004	3.0	0.0053	0.0099

Note: Labyrinth seal: 4-tooth labyrinth; 0.11 in. pitch; 0.11 in. knife height. Brush seal: 0.028 in. brush width; 0.0028-in.-diameter bristles; 0.06 in. fence clearance. Static, 0 rpm.

$$\text{Flow parameter, } \phi = \frac{m\sqrt{T_i}}{P_i A}$$

^aRef. 54.

^bRef. 63.

Effects of Speed. Proctor and Delgado studied the effects of speed (up to 1200 ft/s), temperature (up to 1200°F), and pressure (up to 75 psid) on brush seal and finger seal leakage and power loss.⁶⁴ They determined that leakage generally decreased with increasing speed. It is believed that leakage decreases with speed since the rotor diameter increases, causing both a decrease in the effective seal clearance and an increase in contact stresses.

Aircraft Turbine Engine Performance. Mahler and Boyes⁶⁵ have made leakage comparisons of new and aircraft engine-tested brush seals. They concluded that performance did not deteriorate significantly for periods approaching one engine overhaul cycle (3000 h). Of the three brush seals examined, the “worst-case” brush seal’s leakage rates doubled compared to a new brush seal. Even so, brush seal leakage was still less than half the leakage of the labyrinth seal.

Cryogenic Brush Seals. The long life and low leakage of brush seals make them candidates for use in rocket engine turbopumps. Brush seals 2 in. in diameter with nominal 0.005 in. radial interference were tested in liquid nitrogen (LN₂) at shaft speeds up to 35,000 and 65,000 rpm, respectively, and at pressure drops up to 175 psid per brush.⁶⁶ A labyrinth seal was also tested in liquid nitrogen to provide a baseline. The LN₂ leakage rate of a single brush seal with an initial radial shaft interference of 0.005 in. measured one-half to one-third the leakage rate of a 12-tooth labyrinth seal with a radial clearance of 0.005 in.

Brush seals are not a solution for all seal problems. However, when they are applied within design limits, brush seal leakage will be lower than that of competing labyrinth seals and remain closer to design goals even after transient rub conditions.

Brush Seal Flow Modeling

Brush seal flow modeling is complicated by several factors unique to porous structures in that the leakage depends on the seal porosity, which depends on the pressure drop across the seal. Flow through the brush travels perpendicular to the brush pack through the annulus formed

by the backing ring inner diameter and the shaft diameter, radially inward at successive layers within the brush, and between the bristle tips and the shaft.

A flow model proposed by Holle et al.⁶⁷ uses a single parameter, effective brush thickness, to correlate the flows through the seal. Variation in seal porosity with pressure difference is accounted for by normalizing the varying brush thicknesses by a minimum or ideal brush thickness. Maximum seal flow rates are computed by using an iterative procedure that has converged when the difference in successive iterations for the flow rate is less than a preset tolerance.

Flow models proposed by Hendricks et al.^{68,69} are based on a bulk average flow through the porous media. These models account for brush porosity, bristle loading and deformation, brush geometry parameters, and multiple flow paths. Flow through a brush configuration is simulated by using an electrical analog that has driving potential (pressure drops), current (mass flow), and resistance (flow losses, friction, and momentum) as the key variables. All of the above models require some empirical data to establish the correlating constants. Once these are established the models predict seal flow reasonably well.

A number of researchers (e.g., see Refs. 70–73) have applied numerical techniques to model brush seal flows and bristle pressure loadings. Though these models are more complex, they permit a more detailed investigation of the subtleties of flow and stresses within the brush pack.

Brush Seal Materials

Brush wire bristles range in diameter from 0.0028 in. for low pressures to 0.006 in. for high pressures. The most commonly used material for brush seals is the cobalt-based alloy Haynes 25. Brush seals are generally run against a smooth, hard-face coating to minimize shaft wear and minimize chances of wear-induced cracks from affecting the structural integrity of the rotor. The usual coatings selected are ceramic, including chromium carbide and aluminum oxide. Selecting the correct mating wire and shaft surface finish for a given application can reduce friction heating and extend seal life through reduced oxidation and wear. For extreme operating temperatures to above 1300°F, Derby⁷⁴ has shown low wear and friction for the nickel-based superalloy Haynes 214 (heat treated for high strength) running against a solid-film lubricated hard-face coating Triboglide. Fellenstein et al.^{75,76} investigated a number of bristle/rotor coating material pairs corroborating the benefits of Haynes 25 wires run against chrome carbide but observed Haynes 214 bristle flaring when run against chrome carbide and zirconia coatings.

Nonmetallic Bristles. High-speed turbine designers have long wondered if brush seals could replace labyrinth seals in bearing sump locations. Brush seals would mitigate traditional labyrinth seal clearance opening and corresponding increased leakage. Issues slowing early application of brush seals in these locations included coking (carburization of oil particles at excessively high temperatures), metal particle damage of precision rolling-element bearings, and potential for fires. GE-Global research has found success in applying aramid bristles for certain bearing sump locations.^{77,78} Advantages of the aramid bristles include stable properties up to 300°F (150°C) operating temperatures, negligible amount of shrinkage and moisture absorption, lower wear than Haynes 25 up to 300°F, lower leakage (due to smaller 12- μ m diameters), and resistance to coking.⁷⁷ Based on laboratory demonstration, the aramid fiber seals were installed in a GE 7EA frame (#1) inlet bearing sealing location. Preliminary field data showed that the nonmetallic brush seal maintained a higher pressure difference between the air and bearing drain cavities and enhanced the effectiveness of the sealing system, allowing less oil particles to migrate out of the bearing.

Aircraft Turbine Engines

Applications and Benefits. Brush seals are seeing extensive service in both commercial and military turbine engines. Lower leakage brush seals permit better management of cavity flows and significant reductions in specific fuel consumption when compared to competing labyrinth seals. Allison Engines has implemented brush seals in engines for the Saab 2000, Cessna Citation-X, and V-22 Osprey. GE has implemented a number of brush seals in the balance piston region of the GE90 engine for the Boeing 777 aircraft. PW has entered revenue service with brush seals in three locations⁶⁵ on the PW4168 for Airbus aircraft and on the PW4084 for the Boeing 777.

Ground-Based Turbine Engines. Brush seals are being retrofitted into ground-based turbines both individually and combined with labyrinth seals to greatly improve turbine power output and heat rate (see Refs. 60 and 79–84). Dinc et al.⁶⁰ report that incorporating brush seals in a GE Frame 7EA turbine in the high-pressure packing location increased output by 1.0% and decreased heat rate by 0.5%. Using brush seals in the interstage location resulted in similar improvements. Brush seals have proven effective for service lives of up to 40,000 h!⁶⁰

3.8 Ongoing Developments

Long life and durability under very high temperature ($\geq 1300^\circ\text{F}$) conditions are hurdles to overcome to meet goals of advanced turbine engines under development for next-generation commercial subsonic, supersonic, and military fighter engine requirements. The tribology phenomena are complex and installation specific. In order to extend engine life and bring down maintenance costs, research and development are continuing in this area. To extend brush seal lives at high temperature, Addy et al.,⁶³ Hendricks et al.,⁸⁵ and Howe⁸⁶ have investigated approaches to replace metallic bristles with ceramic fibers. Ceramic fibers offer the potential for operating above 815°C (1500°F) and for reducing bristle wear rates and increasing seal lives while maintaining good flow resistance. Though early results indicate rotor coating wear, ceramic brush leakage rates were less than half those of labyrinth seal (0.007-in. clearance) and bristle wear was low.⁶³

Designers continue to pursue seal designs that address the wear of brush seals. A sample of some of the seal designs being pursued includes the following. Justak has developed a hybrid floating brush seal that combines hydrodynamic seal shoes with a secondary brush seal.^{87,88} Gail and Klemens⁸⁹ have patented a seal that combines a brush with a slide ring to improve sealing effect and reduce wear. Proctor and Steinetz⁹⁰ and Braun et al.⁹¹ are developing an innovative noncontacting finger seal. In this seal, the brush bristles are replaced with precision-machined upstream/downstream finger laminates. Shaft movement is accommodated by bending of the finger elements. Noncontact operation is afforded by hydrodynamic or hydrostatic lift pads located on the downstream laminate. These pads cause the seal to lift during shaft transients. Grondahl has patented a pressure-actuated leaf seal that is designed to overcome wear during transient startup/shutdown conditions.⁹²

REFERENCES

1. I. E. Etsion and B. M. Steinetz "Seals," in H. A. Rothbart (Ed.), *Mechanical Design Handbook*, McGraw-Hill, New York, 1996, Section 17.
2. R. V. Brink, D. E. Czernik, and L. A. Horve, *Handbook of Fluid Sealing*, McGraw-Hill, New York, 1993.
3. A. Bazergui and L. Marchand, "Development of Tightness Test Procedures for Gaskets in Elevated Temperature Service," *Welding Res. Council Bull.*, 339, December 1988.

4. M. Derenne, L. Marchand, J. R. Payne, and A. Bazergui, "Elevated Temperature Testing of Gaskets for Bolted Flanged Connections," *Welding Res. Council Bull.*, 391, May 1994.
5. *Parker O-Ring Handbook*, © 2007 Parker Hannifin Corporation, Cleveland, OH, 2001.
6. L. J. Martini, *Practical Seal Design*, Marcel Dekker, New York, 1984.
7. A. Mathews and G. R. McKillop, "Compression Packings," in *Machine Design Seals Reference Issue*, Penton, March 1967, Chap. 8.
8. R. A. Howard, P. S. Petrunich, and K. C. Schmidt, *Grafoil Engine Design Manual*, Vol. 1, Union Carbide Corp., 1987.
9. B. M. Steinetz and M. L. Adams, "Effects of Compression, Staging and Braid Angle on Braided Rope Seal Performance," *J. Propulsion and Power*, **14**(6), 934–940, December 1998. See also 33rd AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, AIAA-97-2872, Seattle, WA, July 1997, NASA TM, Vol. 107504, July 1997.
10. B. M. Steinetz et al., *High Temperature Braided Rope Seals for Static Sealing Applications*, NASA TM-107233; also *AIAA J. Propulsion and Power*, **13**(5), 1997.
11. E. J. Opila, J. A. Lorincz, and J. J. DeMange, "Oxidation of High-Temperature Alloy Wires in Dry Oxygen and Water Vapor," in E. Opila, J. Fergus, T. Maruyama, J. Mizusaki, T. Narita, D. Shifler, and E. Wuchina (Eds.), *High Temperature Corrosion and Materials Chemistry*, Vol. 5, Electro-Chemical Society, Pennington, NJ, 2005.
12. P. Bauer, *Development of an Enhanced Thermal Barrier for RSRM Nozzle Joints*, AIAA-2000-3566, July 2000.
13. P. Bauer, *MNASA as a Test Bed for Carbon Fiber Thermal Barrier Development*, AIAA-2001-3454, July 2001.
14. P. Totman, A. Prince, D. Frost, and P. Himebaugh, *Alternatives to Silicon Rubber Thermal Barrier in RSRM Nozzle Joints*, AIAA-99-2796, July 1999.
15. M. Ewing, J. R. McGuire, B. B. McWhorter, and D. L. Frost, *Performance Enhancement of the Space Shuttle RSRM Nozzle-to-Case Joint Using a Carbon Rope Barrier*, AIAA-99-2899, July 1999.
16. B. M. Steinetz, and P. H. Dunlap, "Development of Thermal Barriers for Solid Rocket Motor Nozzle Joints," *J. Propulsion Power*, **17**(5), 1023–1034, September/October 2001; also NASA TM-209278, June 1999.
17. B. M. Steinetz and P. H. Dunlap, "Feasibility Assessment of Thermal Barrier Seals for Extreme Transient Temperatures," *J. Propulsion Power*, **16**(2), 347–356, March/April 2000; also NASA TM-208484, July 1998.
18. B. M. Steinetz and P. H. Dunlap, "Rocket Motor Joint Construction Including Thermal Barrier," U.S. Patent No. 6,446,979 B1, September 10, 2002.
19. A. G. Fern and B. S. Nau, *Seals*, Engineering Design Guide 15, published for Design Council, British Standards Institution and Council of Engineering Institutions, Oxford University Press, Oxford, 1976.
20. B. M. Steinetz and R. C. Hendricks, "Aircraft Engine Seals," in *Tribology for Aerospace Applications*, Special Publication SP-37, STLE, 1997, Chapter 9.
- 20a. J. Crane, "Dry Running Noncontacting Gas Seal," Bulletin No. S-3030, 1993.
21. H. Buchter, *Industrial Sealing Technology*, Wiley, New York, 1979.
22. A. O. Lebeck, *Principles and Design of Mechanical Face Seals*, Wiley, New York, 1991.
23. J. Zuk and P. J. Smith, *Quasi-One Dimensional Compressible Flow across Face Seals and Narrow Slots—II. Computer Program*, NASA TN D-6787, 1972.
24. W. F. Hughes et al., *Dynamics of Face and Annular Seals with Two-Phase Flow*, NASA CR-4256, 1989.
25. P. F. Brown, "Status of Understanding for Seal Materials," *Tribology in the 80's*, NASA CP-23000, Vol. 2, 1984, pp. 811–829.
26. J. C. Dahlheimer, *Mechanical Face Seal Handbook*, Chilton Book Co., Philadelphia, 1972.
27. *Mechanical Seal Handbook*, Fluid Sealing Association, Wayne, PA, 2000.
28. *Pumps and Systems Handbook*, Cahaba Media Group, Tuscaloosa, AL, 2003.
29. *Guidelines for Meeting Emission Regulations for Rotating Machinery with Mechanical Seals*, Special Publication SP-30, Society of Tribologists and Lubrication Engineers, Park Ridge, IL, revised 1994.

30. R. C. Waterbury, "Zero-Leak Seals Cut Emissions," *Pumps and Systems Magazine*, AES Marketing, Fort Collins, CO, July 1996.
31. P. E. Bowden, "Design and Selection of Mechanical Seals to Minimize Emissions," *Proc. Inst. Mech. Eng.*, **213**(Pt. J), 1999.
32. H. Hwang, T. Tseng, and B. Shucktis, *Advanced Seals for Engine Secondary Flowpath*, AIAA-95-2618, presented at the 1995 AIAA/ASME/SAE/ASEE Joint Propulsion Conference, San Diego, CA, 1995.
33. C. E. Wolfe et al., "Full Scale Testing and Analytical Validation of an Aspirating Face Seal," AIAA Paper 96-2802, 1996.
34. B. Bagepalli et al., "Dynamic Analysis of an Aspirating Face Seal for Aircraft-Engine Applications," AIAA Paper 96-2803, 1996.
35. J. Munson, "Testing of a High Performance Compressor Discharge Seal," AIAA Paper 93-1997, 1993.
36. R. C. Hendricks, *Seals Code Development—'95*, NASA CP-10181, 1995.
37. Open Channel Software, Chicago, IL.
38. W. Shapiro, *Numerical, Analytical, Experimental Study of Fluid Dynamic Forces in Seals*, Vol. 2: *Description of Gas Seal Codes GCYLT and GFACE*, NASA Contract Report for Contract NAS3-25644, September 1995.
39. J. Walowit and W. Shapiro, *Numerical, Analytical, Experimental Study of Fluid Dynamic Forces in Seals*, Vol. 3: *Description of Spiral-Groove Codes SPIRALG and SPIRALI*, NASA Contract Report for Contract NAS3-25644, September 1995.
40. W. Shapiro et al., *Numerical, Analytical, Experimental Study of Fluid Dynamic Forces in Seals*, Vol. 5: *Description of Seal Dynamics Code DYSEAL and Labyrinth Seals Code KTK*, NASA Contract Report for Contract NAS3-25644, September 1995.
41. R. E. Burcham and R. B. Keller, Jr., *Liquid Rocket Engine Turbopump Rotating-Shaft Seals*, NASA SP-8121, 1979.
42. J. G. Ferguson, "Brushes as High Performance Gas Turbine Seals," ASME Paper 88-GT-182, 1988.
43. A. Egli, "The Leakage of Steam through Labyrinth Seals," *ASME Trans.*, **57**(3), 115–122, 1935.
44. G. Vermes, "A Fluid Mechanics Approach to the Labyrinth Seal Leakage Problem," *J. Eng. Power*, **83**(2), 161–169, 1961.
45. F. H. Mahler, "Advanced Seal Technology," Report PWA-4372, Contract F33615-71-C-1534, Pratt and Whitney Aircraft Co., East Hartford, CT, 1972.
46. R. C. Hendricks, L. T. Tam, and A. Muszynska, *Turbomachine Sealing and Secondary Flows, Part 2—Review of Rotordynamic Issues in Inherently Unsteady Flow Systems with Small Clearances*, NASA TM-2004-211,991, July 2004.
47. D. E. Bently, C. T. Hatch, and B. Grissom (Eds.), *Fundamentals of Rotating Machinery Diagnostics*, Bently Pressurized Bearings, Minden, NV, 2002.
48. J. S. Alford, "Protection of Labyrinth Seals from Flexural Vibration," ASME Paper 63-AHGT-9, 1963; also *J. Eng. Power*, pp. 141–148, April 1964.
49. J. S. Alford, "Protecting Turbomachinery from Self-Excited Rotor Whirl," *J. Eng. Power, Series A*, **87**, 333–344, October 1965.
50. J. S. Alford, "Protecting Turbomachinery from Unstable and Oscillatory Flows," *J. Eng. Power, Series A*, **89**, 513–528, October 1967.
51. H. Benckert and J. Wachter, "Studies on Vibrations Stimulated by Lateral Forces in Sealing Gaps," paper presented at the AGARD Power, Energetics, and Propulsion Meeting on Seal Technology in Gas Turbine Engines, AGARD CP-237 (AGARD AR-123), Paper 9, 1978.
52. D. L. Tipton, T. E. Scott, and R. E. Vogel, *Labyrinth Seal Analysis*, Vol. III: *Analytical and Experimental Development of a Design Model for Labyrinth Seals*, AFWAL TR-85-2103, Allison Gas Turbine Division, General Motors, Indianapolis, IN, 1986.
53. D. L. Rhode and G. H. Nail, "Computation of Cavity-by-Cavity Flow Development in Generic Labyrinth Seals," *J. Tribol.* **14**, 47–51, 1992.
54. H. L. Stocker, D. M. Cox, and G. F. Holle, *Aerodynamic Performance of Conventional and Advanced Design Labyrinth Seals with Solid Smooth, Abradable, and Honeycomb Lands—Gas Turbine Engines*, NASA CR-135307, 1977.

55. Z. Galel, F. Brindisi, and D. Norstrom, "Chemical Stripping of Honeycomb Airseals, Overview and Update," ASME Paper 90-GT-318, 1990.
56. D. W. Childs, D. Elrod, and K. Hale, "Annular Honeycomb Seals: Test Results for Leakage and Rotor-dynamic Coefficients—Comparison to Labyrinth and Smooth Configurations," in *Rotordynamic Instability Problems in High-Performance Turbomachinery*, NASA CP-3026, 1989, pp. 143–159.
57. Perkin Elmer Fluid Sciences product literature, available: <http://fluidsciences.perkinelmer.com/turbomachinery>.
58. Cross Manufacturing product literature, available: www.crossmanufacturing.com.
59. G. F. Holle and M. R. Krishnan, "Gas Turbine Engine Brush Seal Applications," AIAA Paper 90-2142, 1990.
60. S. Dinc, M. Demiroglu, N. Turnquist, G. Toetze, J. Maupin, J. Hopkins, C. Wolfe, and M. Florin, "Fundamental Design Issues of Brush Seals for Industrial Applications," *J. Turbomachinery*, 124, April 2002.
61. J. F. Short et al., "Advanced Brush Seal Development," AIAA Paper 96-2907, 1996.
62. P. Basu et al., "Hysteresis and Bristle Stiffening Effects of Conventional Brush Seals," AIAA Paper 93-1996, 1993.
63. H. E. Addy et al., "Preliminary Results of Silicon Carbide Brush Seal Testing at NASA Lewis Research Center," AIAA Paper 95-2763, 1995.
64. M. P. Proctor and I. R. Delgado, "Leakage and Power Loss Test Results for Competing Turbine Engine Seals," GT2004-53935, in *Proceedings of ASME Turbo Expo, Power for Land, Sea, and Air*, Vienna, Austria, June 2004.
65. F. Mahler and E. Boyes, "The Application of Brush Seals in Large Commercial Jet Engines," AIAA Paper 95-2617, 1995.
66. M. P. Proctor, J. F. Walker, H. D. Perkins, J. F. Hoopes, and G. S. Williamson, "Brush Seals for Cryogenic Applications: Performance, Stage Effects, and Preliminary Wear Results in LN₂ and LH₂," NASA Technical Paper 3536, October 1996.
67. G. F. Holle, R. E. Chupp, and C. A. Dowler, "Brush Seal Leakage Correlations Based on Effective Thickness," paper presented at the Fourth International Symposium on Transport Phenomena and Dynamics of Rotating Machinery, preprint Vol. A., 1992, pp. 296–304.
68. R. C. Hendricks et al., "A Bulk Flow Model of a Brush Seal System," ASME Paper 91-GT-325, 1991.
69. R. C. Hendricks et al., "Investigation of Flows in Bristle and Fiberglass Brush Seal Configurations," paper presented at the Fourth International Symposium on Transport Phenomena and Dynamics of Rotating Machinery, preprint Vol. A, 1992, pp. 315–325.
70. M. J. Braun and V. V. Kudriavtsev, "A Numerical Simulation of Brush Seal Section and Some Experimental Results," *J. Turbomachinery*, 1995.
71. M. T. Turner, J. W. Chew, and C. A. Long, "Experimental Investigation and Mathematical Modeling of Clearance Brush Seals," ASME 97-GT-282, paper presented at the International Gas Turbine and Aeroengine Congress and Exhibition, Orlando, FL, 1997.
72. L. H. Chen, P. E. Wood, T. V. Jones, and J. W. Chew, "An Iterative CFD and Mechanical Brush Seal Model and Comparisons with Experimental Results," ASME 98-GT-372, paper presented at the International Gas Turbine and Aeroengine Congress and Exhibition, Stockholm, Sweden, 1998.
73. M. F. Aksit, "Analysis of Brush Seal Bristle Stresses with Pressure Friction Coupling," ASME GT2003-38718, paper presented at the ASME/IGTI Turbo Expo, Atlanta, GA, June 2003.
74. J. Derby and R. England, "Tribopair Evaluation of Brush Seal Applications," AIAA Paper 92-3715, 1992.
75. J. Fellenstein, C. Della Corte, K. D. Moore, and E. Boyes, *High Temperature Brush Seal Tuft Testing of Metallic Bristles vs Chrome Carbide*, NASA TM-107238, AIAA-96-2908, 1996.
76. J. A. Fellenstein, C. Della Corte, K. A. Moore, and E. Boyes, *High Temperature Brush Seal Tuft Testing of Selected Nickel–Chrome and Cobalt–Chrome Superalloys*, NASA TM-107497, AIAA-97-2634, 1997.
77. N. Bhate, A. C. Thermos, M. F. Aksit, M. Demiroglu, and H. Kizil, "Non-Metallic Brush Seals for Gas Turbine Bearings," GT2004-54296, in *Proceedings of ASME Turbo Expo, Power for Land, Sea, and Air*, Vienna, Austria, June 2004.

78. M. F. Aksit, Y. Dogu, and M. Gursoy, "Hydrodynamic Lift of Brush Seals in Oil Sealing Applications," AIAA-2004-3721, paper presented at the 40th AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Ft. Lauderdale, FL, July 2004.
79. R. E. Chupp, R. P. Johnson, and R. G. Loewenthal, "Brush Seal Development for Large Industrial Gas Turbines," AIAA Paper 95-3146, 1995.
80. R. E. Chupp, R. J. Prior, and R. G. Loewenthal, "Update on Brush Seal Development for Large Industrial Gas Turbines," AIAA Paper 96-3306, 1996.
81. R. E. Chupp, M. F. Aksit, F. Ghasripoor, N. A. Turnquist, and M. Demiroglu, "Advanced Seals for Industrial Turbine Applications," AIAA Paper 2001-3626, 2001.
82. E. Bancalari, I. S. Diakunchak, and G. McQuiggan, "A Review of W501G Engine Design, Development and Field Operating Experience," GT2003-38843, in *Proceedings of ASME Turbo Expo, Power for Land, Sea, and Air*, Atlanta, GA, June 2003.
83. I. S. Diakunchak, G. R. Gaul, G. McQuiggan, and L. R. Southall, "Siemens Westinghouse Advanced Turbine Systems Program Final Summary," GT2002-30654, in *Proceedings of ASME Turbo Expo, Power for Land, Sea, and Air*, Amsterdam, The Netherlands, June 2002.
84. S. Ingistov, "Compressor Discharge Brush Seal for Gas Turbine Model 7EA," *ASME J. of Turbomachinery*, 124, April 2002.
85. R. C. Hendricks, R. Flower, and H. Howe, "Development of a Brush Seals Program Leading to Ceramic Brush Seals," in *Seals Flow Code Development—'93*, NASA CP-10136, 1994, pp. 99–117.
86. H. Howe, "Ceramic Brush Seals Development," in *Seals Flow Code Development—'93*, NASA CP-10136, pp. 133–150, 1994.
87. J. Justak, "Robust Hydrodynamic Brush Seal," U.S. Patent No. 6,428,009, 2002.
88. A. Delgado, L. S. Andres, and J. Justak, "Analysis of Performance and Rotordynamic Force Coefficients of Brush Seals with Reverse Rotation Ability," ASME GT 2004-53614, paper presented at the ASME Turbo Expo 2004 Power for Land, Seal and Air, Vienna, Austria, June 2004.
89. A. Gail and W. Klemens, "Brush Seal," U.S. Patent No. 6,695,314, 2004.
90. M. P. Proctor and B. M. Steinetz, "Non-Contacting Finger Seal," U.S. Patent No. 6,811,154, 2004.
91. M. J. Braun, H. Pierson, D. Deng, and F. Choi, "Non-Contacting Finger Seal Investigations," in *Conference Proceedings of the NASA Seal/Secondary Air Flow System Workshop*, Cleveland, OH, November 2004.
92. C. M. Grondahl, "Seal Assembly and Rotary Machine Containing Such Seal," U.S. Patent No. 6,644,667, 2003.

BIBLIOGRAPHY

- American Society of Mechanical Engineers (ASME), *Code for Pressure Vessels*, Section VIII, Division 1, Appendix 2, ASME, New York, 2004.
- American Variseal, *Variseal™ Design Guide*, AVDG394, American Variseal Co., Broomfield, CO, 2005.
- R. A. Howard, *Grafoil Engineering Design Manual*, Union Carbide, Cleveland, OH, 1987.

CHAPTER 10

STATISTICAL QUALITY CONTROL

Magd E. Zohdi
Louisiana State University
Baton Rouge, Louisiana

1	MEASUREMENTS AND QUALITY CONTROL	325	7	ACCEPTANCE SAMPLING	335
2	DIMENSION AND TOLERANCE	325	7.1	Double Sampling	335
3	QUALITY CONTROL	326	7.2	Multiple and Sequential Sampling	336
3.1	\bar{X} , R , and σ Charts	326	8	DEFENSE DEPARTMENT ACCEPTANCE SAMPLING BY VARIABLES	336
4	INTERRELATIONSHIP OF TOLERANCES OF ASSEMBLED PRODUCTS	331		BIBLIOGRAPHY	336
5	OPERATION CURVE	332			
6	CONTROL CHARTS FOR ATTRIBUTES	332			
6.1	The p and np Charts	333			
6.2	The c and u Charts	334			

1 MEASUREMENTS AND QUALITY CONTROL

The metric and English measuring systems are the two measuring systems commonly used throughout the world. The metric system is universally used in most scientific applications, but, for manufacturing in the United States it has been limited to a few specialties, mostly items that are related in some way to products manufactured abroad.

2 DIMENSION AND TOLERANCE

In dimensioning a drawing, the numbers placed in the dimension lines are only approximate and do not represent any degree of accuracy unless so stated by the designer. To specify the degree of accuracy, it is necessary to add tolerance figures to the dimension. Tolerance is the amount of variation permitted in the part or the total variation allowed in a given dimension.

Dimensions given close tolerances mean that the part must fit properly with some other part. Both must be given tolerances in keeping with the allowance desired, the manufacturing processes available, and the minimum cost of production and assembly that will maximize profit. Generally speaking, the cost of a part goes up as the tolerance is decreased.

Allowance, which is sometimes confused with tolerance, has an altogether different meaning. It is the minimum clearance space intended between mating parts and represents the condition of tightest permissible fit.

3 QUALITY CONTROL

When parts must be inspected in large numbers, 100% inspection of each part is not only slow and costly but does not eliminate all of the defective pieces. Mass inspection tends to be careless, operators become fatigued, and inspection gauges become worn or out of adjustment more frequently. The risk of passing defective parts is variable and of unknown magnitude, whereas, in a planned sampling procedure, the risk can be calculated. Many products, such as bulbs, cannot be 100% inspected since any final test made on one results in the destruction of the product. Inspection is costly and nothing is added to a product that has been produced to specifications.

Quality control enables an inspector to sample the parts being produced in a mathematical manner and to determine whether or not the entire stream of production is acceptable, provided that the company is willing to allow up to a certain known number of defective parts. This number of acceptable defectives is usually taken as 3 out of 1000 parts produced. Other values might be used.

3.1 \bar{X} , R , and σ Charts

To use quality techniques in inspection, the following steps must be taken (see Table 1):

1. Sample the stream of products by taking m samples, each of size n .
2. Measure the desired dimension in the sample, mainly the central tendency.
3. Calculate the deviations of the dimensions.
4. Construct a control chart.
5. Plot succeeding data on the control chart.

The arithmetic mean of the set of n units is the main measure of central tendency. The symbol \bar{X} is used to designate the arithmetic mean of the sample and may be expressed in algebraic terms as

$$\bar{X}_i = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} \quad (1)$$

where X_1, X_2, X_3 , etc. represent the specific dimensions in question. The most useful measure of dispersion of a set of numbers is the standard deviation σ . It is defined as the root-mean-square deviation of the observed numbers from their arithmetic mean. The standard deviation σ is expressed in algebraic terms as

$$\sigma_i = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n}} \quad (2)$$

Table 1 Computational Format for Determining \bar{X} , R , and σ

Sample Number	Sample Values	Mean \bar{X}	Range R	Standard Deviation σ'
1	$X_{11}, X_{12}, \dots, X_{1n}$	\bar{X}_1	R_1	σ'_1
2	$X_{21}, X_{22}, \dots, X_{2n}$	\bar{X}_2	R_2	σ'_2
.
.
.
m	$X_{m1}, X_{m2}, \dots, X_{mn}$	\bar{X}_m	R_m	σ'_m

Another important measure of dispersion, used particularly in control charts, is the range R . The range is the difference between the largest observed value and the smallest observed in a specific sample:

$$R = X_i(\max) - X_i(\min) \quad (3)$$

Even though the distribution of the X values in the universe can be of any shape, the distribution of the \bar{X} values tends to be close to the normal distribution. The larger the sample size and the more nearly normal the universe, the closer will the frequency distribution of the average \bar{X} approach the normal curve, as in Fig. 1.

According to the statistical theory (the central limit theory), in the long run, the average of the \bar{X} values will be the same as μ , the average of the universe. And in the long run, the standard deviation of the frequency distribution \bar{X} values $\sigma_{\bar{x}}$ will be given by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (4)$$

where σ is the standard deviation of the universe. To construct the control limits, the following steps are taken:

1. Calculate the average of the average $\bar{\bar{X}}$ as follows:

$$\bar{\bar{X}} = \sum_1^m m/\bar{X}_i \quad i = 1, 2, \dots, m \quad (5)$$

2. Calculate the average deviation, $\bar{\sigma}$ where

$$\bar{\sigma} = \sum_1^m m/\sigma'_i \quad i = 1, 2, \dots, m \quad (6)$$

Statistical theory predicts the relationship between $\bar{\sigma}$ and $\sigma_{\bar{x}}$. The relationship for the $3\sigma_{\bar{x}}$ limits or the 99.73% limits is

$$A_1 \bar{\sigma} = 3\sigma_{\bar{x}} \quad (7)$$

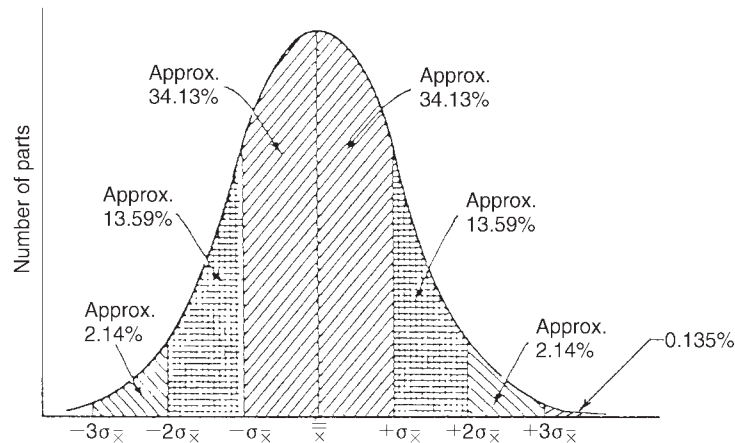


Figure 1 Normal distribution and percentage of parts that will fall within σ limits.

Table 2 Factors for \bar{X} , R , σ , and X Control Charts

Sample Size n	Factors for \bar{X} Chart		Factors for R Chart		Factors for σ Chart		Factors for X Chart		$\frac{\sigma = \bar{R}/d_2}{d_2}$
	From \bar{R} A_2	From $\bar{\sigma}$ A_1	Lower D_3	Upper D_4	Lower B_3	Upper B_4	From \bar{R} E_2	From $\bar{\sigma}$ E_1	
2	1.880	3.759	0	3.268	0	3.267	2.660	5.318	1.128
3	1.023	2.394	0	2.574	0	2.568	1.772	4.146	1.693
4	0.729	1.880	0	2.282	0	2.266	1.457	3.760	2.059
5	0.577	1.596	0	2.114	0	2.089	1.290	3.568	2.326
6	0.483	1.410	0	2.004	0.030	1.970	1.184	3.454	2.539
7	0.419	1.277	0.076	1.924	0.118	1.882	1.109	3.378	2.704
8	0.373	1.175	0.136	1.864	0.185	1.815	1.054	3.323	2.847
9	0.337	1.094	0.184	1.816	0.239	1.761	1.011	3.283	2.970
10	0.308	1.028	0.223	1.777	0.284	1.716	0.975	3.251	3.078
11	0.285	0.973	0.256	1.744	0.321	1.679	0.946	3.226	3.173
12	0.266	0.925	0.284	1.717	0.354	1.646	0.921	3.205	3.258
13	0.249	0.884	0.308	1.692	0.382	1.618	0.899	3.188	3.336
14	0.235	0.848	0.329	1.671	0.406	1.594	0.881	3.174	3.407
15	0.223	0.817	0.348	1.652	0.428	1.572	0.864	3.161	3.472
16	0.212	0.788	0.364	1.636	0.448	1.552	0.848	3.152	3.532
17	0.203	0.762	0.380	1.621	0.466	1.534	0.830	3.145	3.588
18	0.194	0.738	0.393	1.608	0.482	1.518	0.820	3.137	3.640
19	0.187	0.717	0.404	1.597	0.497	1.503	0.810	3.130	3.687
20	0.180	0.698	0.414	1.586	0.510	1.490	0.805	3.122	3.735
21	0.173	0.680	0.425	1.575	0.523	1.477	0.792	3.114	3.778
22	0.167	0.662	0.434	1.566	0.534	1.466	0.783	3.105	3.819
23	0.162	0.647	0.443	1.557	0.545	1.455	0.776	3.099	3.858
24	0.157	0.632	0.451	1.548	0.555	1.445	0.769	3.096	3.895
25	0.153	0.619	0.459	1.540	0.565	1.435	0.765	3.095	3.931

This means that control limits are set so that only 0.27% of the produced units will fall outside the limits. The value of $3\sigma_{\bar{x}}$ is an arbitrary limit that has found acceptance in industry.

The value of A_1 calculated by probability theory is dependent on the sample size and is given in Table 2. The formula for 3σ control limits (CL) using this factor is

$$CL(\bar{X}) = \bar{\bar{X}} \pm A_1 \bar{\sigma} \tag{8}$$

Once the control chart (Fig. 2) has been established, data (\bar{X}_i 's) that result from samples of the same size n are recorded on it. It becomes a record of the variation of the inspected dimensions over a period of time. The data plotted should fall in random fashion between the control limits 99.73% of the time if a stable pattern of variation exists.

So long as the points fall between the control lines, no adjustments or changes in the process are necessary. If five to seven consecutive points fall on one side of the mean, the process should be checked. When points fall outside of the control lines, the cause must be located and corrected immediately.

Statistical theory also gives the expected relationship between $\bar{R}(\Sigma R_i/m)$ and $\sigma_{\bar{x}}$. The relationship for the $3\sigma_{\bar{x}}$ limits is

$$A_2 \bar{R} = 3\sigma_{\bar{x}} \tag{9}$$

The values for A_2 calculated by probability theory, for different sample sizes, are given in Table 2.

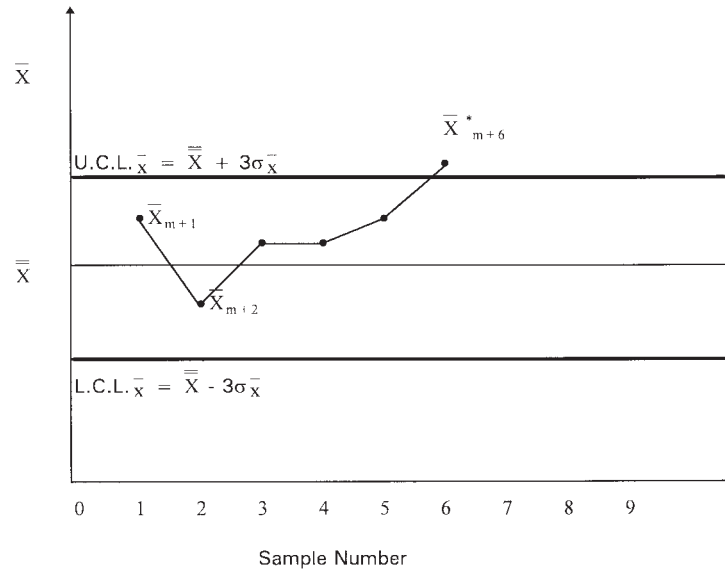


Figure 2 Control chart \bar{X} .

The formula for 3σ control limits using this factor is

$$CL(\bar{X}) = \bar{\bar{X}} \pm A_2 \bar{R} \quad (10)$$

In control chart work, the ease of calculating R is usually much more important than any slight theoretical advantage that might come from the use of σ . However, in some cases where the measurements are costly and it is necessary that the inferences from a limited number of tests be as reliable as possible, the extra cost of calculating σ is justified. It should be noted that, because Fig. 2 shows the averages rather than individual values, it would have been misleading to indicate the tolerance limits on this chart. It is the individual article that has to meet the tolerances, not the average of a sample. Tolerance limits should be compared to the machine capability limits. Capability limits are the limits on a single unit and can be calculated by

$$\begin{aligned} \text{Capability limits} &= \bar{\bar{X}} \pm 3\sigma \\ \sigma &= d_2 / \bar{R} \end{aligned} \quad (11)$$

Since $\sigma' = \sqrt{n} \sigma_{\bar{x}}$ the capability limits can be given by

$$\text{Capability limits (X)} = \bar{\bar{X}} \pm 3\sqrt{n} \sigma_{\bar{x}} \quad (12)$$

$$= \bar{\bar{X}} \pm E_1 \bar{\sigma} \quad (13)$$

$$= \bar{\bar{X}} \pm E_2 \bar{R} \quad (14)$$

The values for d_2 , E_1 , and E_2 calculated by probability theory, for different sample sizes, are given in Table 2.

Figure 3 shows the relationships among the control limits, the capability limits, and assumed tolerance limits for a machine that is capable of producing the product with this

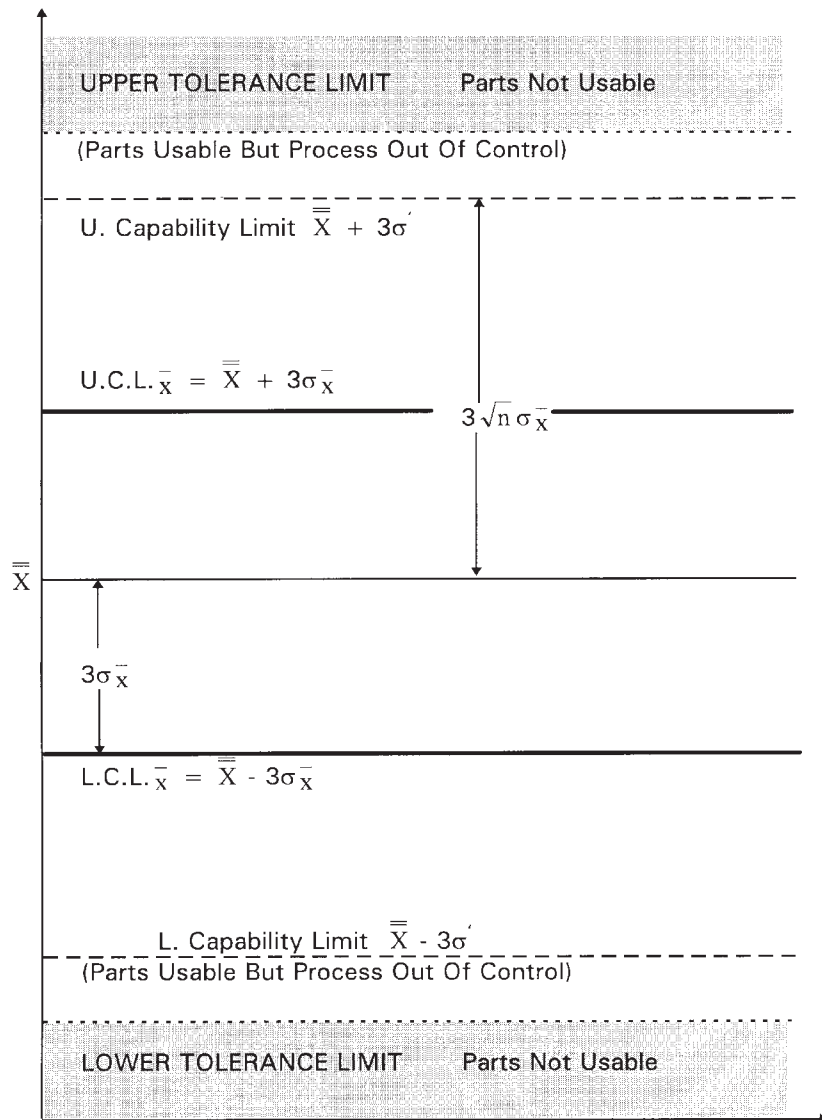


Figure 3 Control, capability, and tolerance (specification limits).

specified tolerance. Capability limits indicate that the production facility can produce 99.73% of its products within these limits. If the specified tolerance limits are greater than the capability limits, the production facility is capable of meeting the production requirement.

If the specified tolerance limits are tighter than the capability limits, a certain percentage of the production will not be usable and 100% inspection will be required to detect the products outside the tolerance limits.

To detect changes in the dispersion of the process, the R and σ charts are often employed with \bar{X} and X charts.

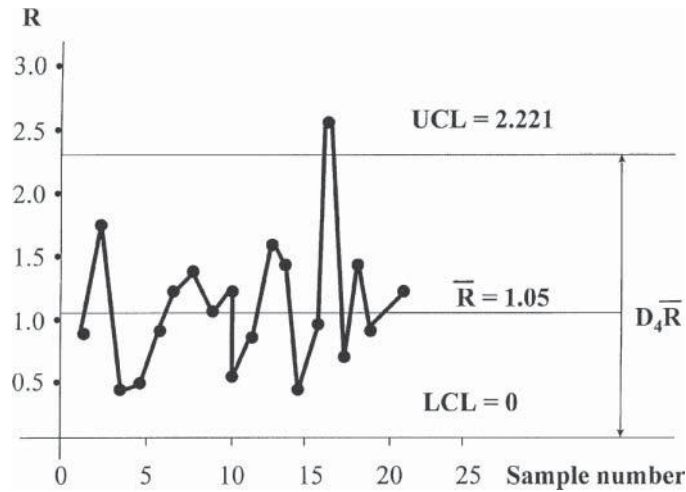


Figure 4 R Chart for samples of 5 each.

The upper (UCL) and lower (LCL) control limits for the R chart are specified as

$$UCL (R) = D_4 \bar{R} \tag{15}$$

$$LCL (R) = D_3 \bar{R} \tag{16}$$

Figure 4 shows the \bar{R} chart for samples of size 5.

The upper and lower control for the T chart are specified as

$$UCL (\sigma) = B_4 \bar{\sigma} \tag{17}$$

$$LCL (\sigma) = B_3 \bar{\sigma} \tag{18}$$

The values for D_3 , D_4 , B_3 , and B_4 calculated by probability theory, for different sample sizes, are given in Table 2.

4 INTERRELATIONSHIP OF TOLERANCES OF ASSEMBLED PRODUCTS

Mathematical statistics states that the dimension on an assembled product may be the sum of the dimensions of the several parts that make up the product. It states also that the standard deviation of the sum of any number of independent variables is the square root of the sum of the squares of the standard deviations of the independent variables. So if

$$X = X_1 \pm X_2 \pm \dots \pm X_n \tag{19}$$

$$\bar{X} = \bar{X}_1 \pm \bar{X}_2 \pm \dots \pm \bar{X}_n \tag{20}$$

$$\sigma(X) = \sqrt{(\sigma_1)^2 + (\sigma_2)^2 + \dots + (\sigma_n)^2} \tag{21}$$

Whenever it is reasonable to assume that the tolerance ranges of the parts are proportional to their respective σ' values, such tolerance ranges may be combined by taking the square root of the sum of the squares:

$$T = \sqrt{T_1^2 + T_2^2 + T_3^2 + \dots + T_n^2} \tag{22}$$

5 OPERATION CURVE

Control charts detect changes in a pattern of variation. If the chart indicates that a change has occurred when it has not, type I error occurs. If 3σ limits are used, the probability of making a type I error is approximately 0.0027.

The probability of the chart indicating no change, when in fact it has, is the probability of making a type II error. The operation characteristic curves are designed to indicate the probability of making a type II error. An operation curve (OC) for an \bar{X} chart of 3σ limits is illustrated in Fig. 5.

6 CONTROL CHARTS FOR ATTRIBUTES

Testing may yield only one of two defined classes: within or outside certain limits, acceptable or defective, working or idle. In such a classification system, the proportion of units falling in one class may be monitored with a p chart.

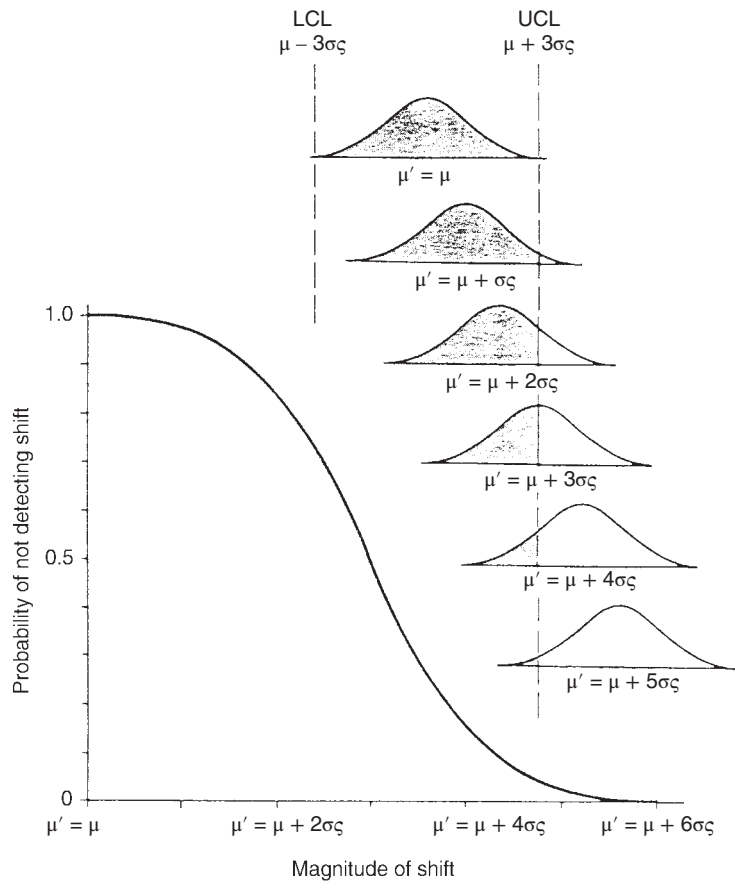


Figure 5 Operating characteristic curve for 3σ limit.

In other cases, observation may yield a multivalued, but still discrete, classification system. In such case, the number of discrete observations, such as events, objects, states, or occurrences, may be monitored by a c chart.

6.1 The p and np Charts

When sampled items are tested and placed into one of two defined classes, the proportion of units falling into one class p is described by the binomial distribution. The mean and standard deviation are given as

$$m = np$$

$$\sigma = \sqrt{np(1-p)}$$

Dividing by the sample size n , the parameters are expressed as proportions. These statistics can be expressed as

$$\bar{p} = \frac{\text{Total number in the class}}{\text{Total number of observations}} \quad (23)$$

$$\sigma_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad (24)$$

The control limits are either set at 2σ limits with type I error as 0.0456 or at 3σ limits with type I error as 0.0027. The control limits for the p chart with 2σ limits (Fig. 6) are defined as

$$CL(p) = \bar{p} \pm 2S_p \quad (25)$$

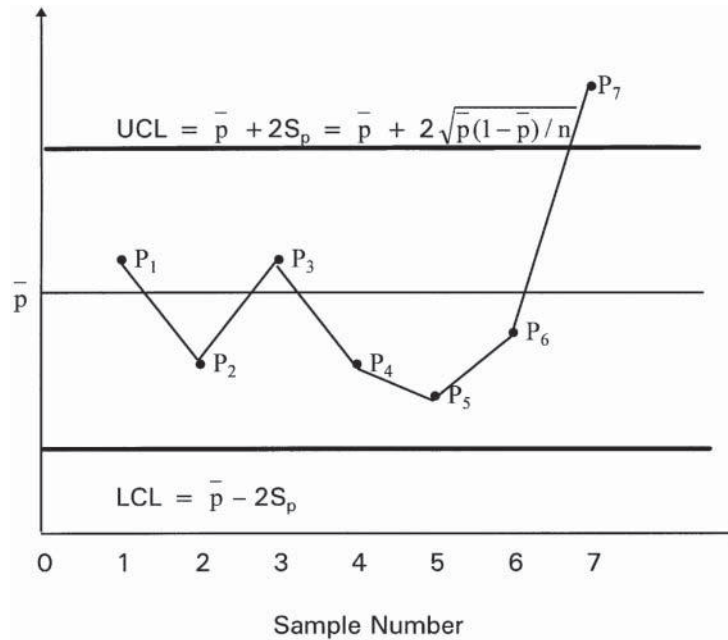


Figure 6 P charts.

However, if subgroup size is constant, the chart for actual numbers of rejects np or pn may be used. The appropriate model for 3σ control limits on an np chart is

$$CL(np) = n\bar{p} \pm 3\sqrt{n\bar{p}(1-\bar{p})} \quad (26)$$

6.2 The c and u Charts

The random variable process that provides numerical data that are recorded as a number c rather than a proportion p is described by the Poisson distribution. The mean and the variance of the Poisson distribution are equal and expressed as $\mu = \sigma^2 = np$. The Poisson distribution is applicable in any situation when n and p cannot be determined separately, but their product np can be established. The mean and variable can be estimated as

$$\bar{c} = S_c^2 = \frac{\sum_{i=1}^m C_i}{m} = \frac{\sum_{i=1}^m (np)_i}{m} \quad (27)$$

The control limits (Fig. 7) are defined as

$$CL(c) = \bar{C} \pm 3S_c \quad (28)$$

If there is change in the area of opportunity for occurrence of a nonconformity from subgroup to subgroup, such as number of units inspected or the lengths of wires checked, the conventional c chart showing only the total number of nonconformities is not applicable. To create some standard measure of the area of opportunity, the nonconformities per unit (c/n) or u is used as

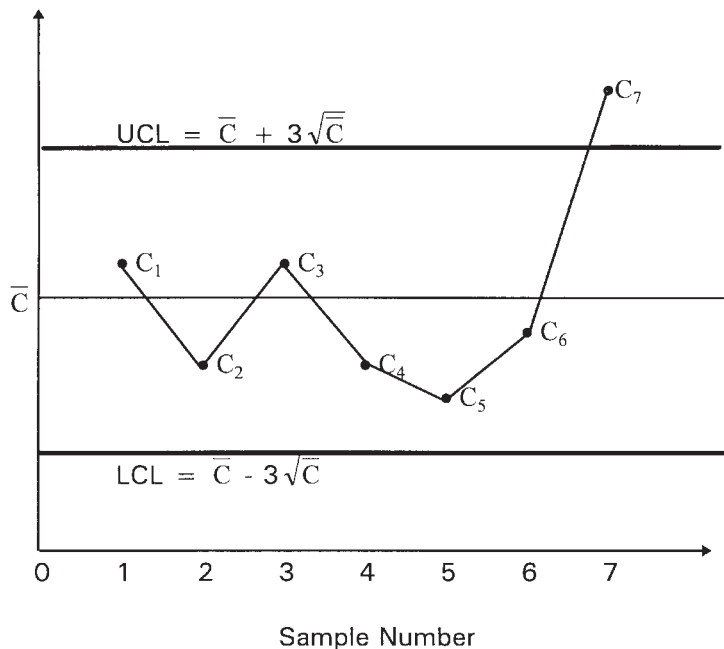


Figure 7 C charts.

the control statistic. The control limits are

$$CL(u)\bar{u} \pm 3 \frac{\sqrt{\bar{u}}}{\sqrt{n_i}} \quad (29)$$

$$\text{where } \bar{u} = \frac{\sum C_i}{\sum n_i} = \frac{\text{Total nonconformities found}}{\text{Total units inspected}}$$

$c = nu$ is Poisson distributed, u is not

7 ACCEPTANCE SAMPLING

The objective of acceptance sampling is to determine whether a quantity of the output of a process is acceptable according to some criterion of quality. A sample from the lot is inspected and the lot is accepted or rejected in accordance with the findings of the sample.

Acceptance sampling plans call for the random selection of sample of size n from a lot containing N items. The lot is accepted if the number of defectives found in the sample are $\leq c$, the acceptance number. A rejected lot can either be returned to the producer, nonrectifying inspections, or it can be retained and subjected to a 100% screening process, rectifying inspection plan improves the outgoing quality. A second attribute inspection plan might use two samples before requiring the acceptance or rejection of a lot. A third plan might use multiple samples or a sequential sampling process in evaluating a lot. Under rectifying inspection programs, the average outgoing quality level (AOQ), the average inspection lot (I), and the average outgoing quality limit (AOQL) can be predicted for varying levels of incoming fraction defective p .

Assuming that all lots arriving contain the same proportion of defectives p , and that rejected lots will be subjected to 100% inspection, AOQ and I are given as:

$$AOQ = \frac{P_a p(N - n)}{N - pn - (1 - P_a)p(Nn)} \quad (30)$$

$$I = n + (1 - P_a)(N - n) \quad (31)$$

The AOQ increases as the proportion defective in incoming lots increases until it reaches a maximum value and then starts to decrease. This maximum value is referred to as the AOQL. The hypergeometric distribution is the appropriate distribution to calculate the probability of acceptance P_a ; however, the Poisson distribution is used as an approximation.

Nonrectifying inspection program does not significantly improve the quality level of the lots inspected.

7.1 Double Sampling

Double sampling involves the possibility of putting off the decision on the lot until a second sample has been taken. A lot may be accepted at once if the first sample is good enough or rejected at once if the first sample is bad enough. If the first sample is neither, the decision is based on the evidence of the first and second samples combined.

The symbols used in double sampling are

N = lot size

n_1 = first sample

c_1 = acceptance number for first sample

n_2 = second sample

c_2 = acceptance number of the two samples combined

Computer programs are used to calculate the OC curves: acceptance after the first sample, rejection after the first sample, acceptance after the second sample, and rejection after the second sample. The average sample number (ASN) in double sampling is given by

$$\text{ASN} = [P_a(n_1) + P_r(n_1)]n_1 + [P_a(n_2) + P_r(n_2)](n_1 + n_2) \quad (32)$$

7.2 Multiple and Sequential Sampling

In multiple sampling, three or more samples of a stated size are permitted and the decision on acceptance or rejection is revealed after a stated number of samples. In sequential sampling, item-by-item inspection, a decision is possible after each item has been inspected and when there is no specified limit on the total number of units to be inspected. Operation curves are developed through computer programs. The advantage of using double sampling, multiple sampling, or sequential sampling is to reach the appropriate decision with fewer items inspected.

8 DEFENSE DEPARTMENT ACCEPTANCE SAMPLING BY VARIABLES

MIL-STD-105 A, B, C, D, and then ABC-STD-105, are based on the acceptance quality level (AQL) concept. The plans contain single, double, or multiple sampling, depending on the lot size and AQL and the probability of acceptance at this level P_a . Criteria for shifting to tightened inspection, requalification for normal inspection, and reduced inspection are listed in the tables associated with the plan.

MIL-STD-414 plans were developed to reduce inspection lots by using sample sizes compared to MIL-STD-105. They are similar, as both procedures and tables are based on the concept of AQL; lot-by-lot acceptance inspection; both provide for normal, tightened, or reduced inspection; sample sizes are greatly influenced by lot size; several inspection levels are available; and all plans are identified by sample size code letter. MIL-STD-414 could be applied either with a single specification limit, L or U , or with two specification limits. Known sigma plans included in the standard were designated as having “variability known.” Unknown sigma plans were designated as having “variability unknown.” In the latter-type plans, it was possible to use either the standard deviation method or the range method in estimating the lot variability.

BIBLIOGRAPHY

- L. S. Aft, *Fundamentals of Industrial Quality Control*, 3rd ed., Addison–Wesley, Menlo Park, CA, 1988.
- ASTM Manual on Presentation of Data and Control Chart Analysis*, Special Technical Pub. 15D, American Society for Testing and Materials, Philadelphia, PA, 1976.
- B. Bhushan (Ed.), *Modern Tribology Handbook*, CRC Press, Boca Raton, FL, 2001.
- R. Clements, *Quality ITQM/ISO 9000*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
- Control Chart Method of Controlling Quality During Production, ANSI Standard 21.3-1975*, American National Standards Institute, New York, 1975.
- J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, Duxburg Press, New York, 1995.
- H. F. Dodge, *A General Procedure for Sampling Inspection by Attributes—Based on the AQL Concept*, Technical Report No. 10, The Statistics Center, Rutgers—The State University, New Brunswick, NJ, 1959.
- P. J. Drake, *Dimension and Tolerancing Handbook*, McGraw-Hill, New York, 1999.
- A. J. Duncan, *Quality Control and Industrial Statistics*, 5th ed., Richard D. Irwin, Homewood, IL, 1986.
- A. V. Feigenbaum, *Total Quality Control—Engineering and Management*, 3rd ed., McGraw-Hill, New York, 1991.
- E. L. Grant, and R. S. Leavenworth, *Statistical Quality Control*, 7th ed., McGraw-Hill, New York, 1996. (Software included)

- J. M. Juran, and F. M. Gryna, Jr., *Quality Control Handbook*, McGraw-Hill, New York, 1988.
- J. M. Juran, and F. M. Gryna, Jr., *Quality Planning and Analysis*, 4th ed., McGraw-Hill, New York, 2000.
- J. Lamprecht, *Implementing the ISO 9000 Series*, Marcel Dekker, New York, 1995.
- Military Standard 105E, Sampling Procedures and Tables for Inspection by Attributes*, Superintendent of Documents, Government Printing Office, Washington, DC, 1969.
- Military Standard 414, Sampling Procedures and Tables for Inspection by Variables for Percent Defective*, Superintendent of Documents, Government Printing Office, Washington, DC, 1957.
- Military Standard 690-B, Failure Rate Sampling Plans and Procedures*, Superintendent of Documents, Government Printing Office, Washington, DC, 1969.
- Military Standard 781-C, Reliability Design Qualification and Production Acceptance Tests, Exponential Distribution*, Superintendent of Documents, Government Printing Office, Washington, DC, 1977.
- Military Standard 1235B, Single and Multi-Level Continuous Sampling Procedures and Tables for Inspection by Attributes*, Superintendent of Documents, Government Printing Office, Washington, DC, 1981.
- D. C. Montgomery, and G. C. Runger, *Introduction to Statistical Quality Control*, Wiley, New York, 2001.
- J. Rabbit, and P. Bergh, *The ISO 9000 Book*, ME, Dearborn, MI, 1993.
- Society of Manufacturing Engineers, *Quality Control and Assembly*, 4th ed., Dearborn, MI, 1994.
- Supply and Logistic Handbook—Inspection H 105. Administration of Sampling Procedures for Acceptance Inspection*, Superintendent of Documents, Government Printing Office, Washington, DC, 1954.
- R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and Statistics for Engineers and Scientists*, Prentice Hall, Englewood Cliffs, NJ, 2002.
- M. E. Zohdi, *Manufacturing Processes Quality Evaluation and Testing*, International Conference, Operations Research, January 1976.

CHAPTER 11

COMPUTER-INTEGRATED MANUFACTURING

William E. Biles
University of Louisville
Louisville, Kentucky

Magd E. Zohdi
Louisiana State University
Baton Rouge, Louisiana

1 INTRODUCTION	339	4.3 Robot Control and Programming	349
2 DEFINITIONS AND CLASSIFICATIONS	340	4.4 Robot Applications	350
2.1 Automation	340	5 COMPUTERS IN MANUFACTURING	351
2.2 Production Operations	341	5.1 Hierarchical Computer Control	351
2.3 Production Plants	342	5.2 CNC and DNC Systems	352
2.4 Models for Production Operations	342	5.3 Manufacturing Cell	352
3 NUMERICAL CONTROL MANUFACTURING SYSTEMS	344	5.4 Flexible Manufacturing Systems	352
3.1 Numerical Control	344	6 GROUP TECHNOLOGY	353
3.2 Coordinate System	345	6.1 Part Family Formation	354
3.3 Selection of Parts for NC Machining	346	6.2 Parts Classification and Coding	354
3.4 CAD/CAM Part Programming	346	6.3 Production Flow Analysis	357
3.5 Programming by Scanning and Digitizing	347	6.4 Types of Machine Cell Designs	357
3.6 Adaptive Control	347	6.5 Computer-Aided Process Planning	358
3.7 Machinability Data Prediction	347	BIBLIOGRAPHY	358
4 INDUSTRIAL ROBOTS	349		
4.1 Definition	349		
4.2 Robot Configurations	349		

1 INTRODUCTION

Modern manufacturing systems are advanced automation systems that use computers as an integral part of their control. Computers are a vital part of automated manufacturing. They control stand-alone manufacturing systems, such as various machine tools, welders, laser beam cutters, robots, and automatic assembly machines. They control production lines and are beginning to take over control of the entire factory. The computer-integrated manufacturing system (CIMS) is a reality in the modern industrial society. As illustrated in Fig. 1, CIMS combines computer-aided design (CAD), computer-aided manufacturing (CAM), computer-aided

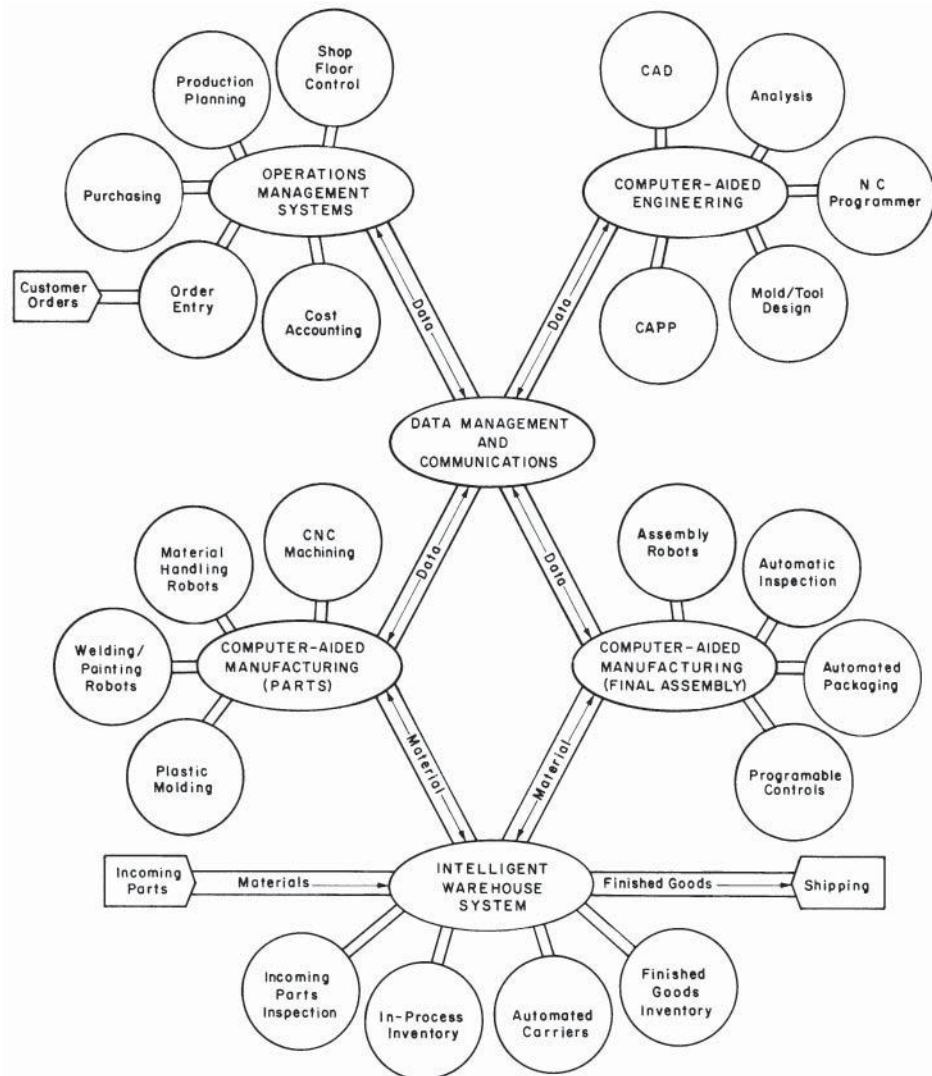


Figure 1 Computer-integrated manufacturing system.

inspection (CAI), and computer-aided production planning (CAPP), along with automated material handling. This chapter focuses on CAM for both parts fabrication and assembly, as shown in Fig. 1. It treats numerical control (NC) machining, robotics, and group technology. It shows how to integrate these functions with automated material storage and handling to form a CIM system.

2 DEFINITIONS AND CLASSIFICATIONS

2.1 Automation

Automation is a relatively new word, having been coined in the 1930s as a substitute for the word *automatization*, which referred to the introduction of automatic controls in manufacturing.

Automation implies the performance of a task without human assistance. Manufacturing processes are classified as manual, semiautomatic, or automatic, depending on the extent of human involvement in the ongoing operation of the process.

The primary reasons for automating a manufacturing process are to:

1. Reduce the cost of the manufactured product through savings in both material and labor
2. Improve the quality of the manufactured product by eliminating errors and reducing the variability in product quality
3. Increase the rate of production
4. Reduce the lead time for the manufactured product, thus providing better service for customers
5. Make the workplace safer

The economic reality of the marketplace has provided the incentive for industry to automate its manufacturing processes. In Japan and in Europe, the shortage of skilled labor sparked the drive toward automation. In the United States, stern competition from Japanese and European manufacturers, in terms of both product cost and product quality, has necessitated automation. Whatever the reasons, a strong movement toward automated manufacturing processes is being witnessed throughout the industrial nations of the world.

2.2 Production Operations

Production is a transformation process in which raw materials are converted into the goods demanded in the marketplace. Labor, machines, tools, and energy are applied to materials at each of a sequence of steps that bring the materials closer to a marketable final state. These individual steps are called *production operations*.

There are three basic types of industries involved in transforming raw materials into marketable products:

1. *Basic Producers*. These transform natural resources into raw materials for use in manufacturing industry—for example, iron ore to steel ingot in a steel mill.
2. *Converters*. These take the output of basic producers and transform the raw materials into various industrial products—for example, steel ingot is converted into sheet metal.
3. *Fabricators*. These fabricate and assemble final products—for example, sheet metal is fabricated into body panels and assembled with other components into an automobile.

The concept of a CIMS as depicted in Fig. 1 applies specifically to a “fabricator” type of industry. It is the “fabricator” industry that we focus on in this chapter.

The steps involved in creating a product are known as the “manufacturing cycle.” In general, the following functions will be performed within a firm engaged in manufacturing a product:

1. *Sales and Marketing*. The order to produce an item stems either from customer orders or from production orders based on product demand forecasts.
2. *Product Design and Engineering*. For proprietary products, the manufacturer is responsible for development and design, including component drawings, specifications, and bill of materials.
3. *Manufacturing Engineering*. Ensuring manufacturability of product designs, process planning, design of tools, jigs, and fixtures and “troubleshooting” the manufacturing process.
4. *Industrial Engineering*. Determining work methods and time standards for each production operation.

5. *Production Planning and Control*. Determining the master production schedule, engaging in material requirements planning, operations scheduling, dispatching job orders, and expediting work schedules.
6. *Manufacturing*. Performing the operations that transform raw materials into finished goods.
7. *Material Handling*. Transporting raw materials, in-process components, and finished goods between operations.
8. *Quality Control*. Ensuring the quality of raw materials, in-process components, and finished goods.
9. *Shipping and Receiving*. Sending shipments of finished goods to customers or accepting shipments of raw materials, parts, and components from suppliers.
10. *Inventory Control*. Maintaining supplies of raw materials, in-process items, and finished goods so as to provide timely availability of these items when needed.

Thus, the task of organizing and coordinating the activities of a company engaged in the manufacturing enterprise is complex. The field of industrial engineering is devoted to such activities.

2.3 Production Plants

There are several ways to classify production facilities. One way is to refer to the volume or rate of production. Another is to refer to the type of plant layout. Actually, these two classification schemes are related, as will be pointed out.

In terms of the volume of production, there are three types of manufacturing plants:

1. *Job Shop Production*. Commonly used to meet specific customer orders; great variety of work; production equipment must be flexible and general purpose; high skill level among workforce—for example, aircraft manufacturing.
2. *Batch Production*. Manufacture of product in medium lot sizes; lots produced only once at regular intervals; general-purpose equipment, with some specialty tooling—for example, household appliances, lawn mowers.
3. *Mass Production*. Continuous specialized manufacture of identical products; high production rates; dedicated equipment; lower labor skills than in a job shop or batch manufacturing—for example, automotive engine blocks.

In terms of the arrangement of production resources, there are three types of plant layouts:

1. *Fixed-Position Layout*. The item is placed in a specific location and labor and equipment are brought to the site. Job shops often employ this type of plant layout.
2. *Process Layout*. Production machines are arranged in groups according to the general type of manufacturing process; forklifts and hand trucks are used to move materials from one work center to the next. Batch production is most often performed in process layouts.
3. *Product flow Layout*. Machines are arranged along a line or in a U or S configuration, with conveyors transporting work parts from one station to the next; the product is progressively fabricated as it flows through the succession of workstations. Mass production is usually conducted in a product flow layout.

2.4 Models for Production Operations

In this section, we examine three types of models by which we can examine production operations: graphical models, manufacturing process models, and mathematical models of

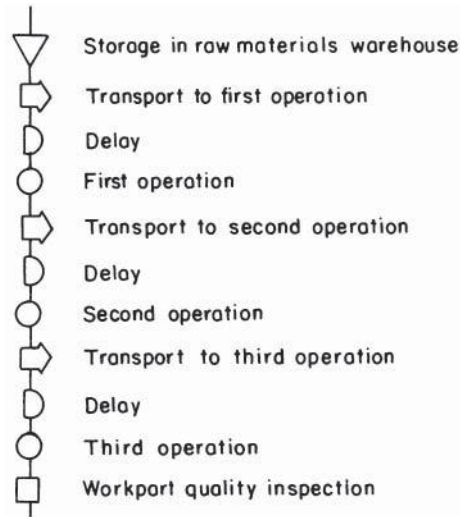


Figure 2 Flow process chart for a sample work part.

production activity. Process flow charts depict the sequence of operations, storages, transportations, inspections, and delays encountered by a work part of assembly during processing. As illustrated in Fig. 2, a process flow chart gives no representation of the layout or physical dimensions of a process but focuses on the succession of steps seen by the product. It is useful in analyzing the efficiency of the process in terms of the proportion of time spent in transformation operations as opposed to transportations, storages, and delays.

The manufacturing process model gives a graphical depiction of the relationship among the several entities that comprise the process. It is an input–output model. Its inputs are raw materials, equipment (machine tools), tooling and fixtures, energy, and labor. Its outputs are completed workpieces, scrap, and waste. These are shown in Fig. 3. Also shown in this figure are the controls that are applied to the process to optimize the utilization of the inputs in producing completed workpieces or in maximizing the production of completed workpieces at a given set of values describing the inputs.

Mathematical models of production activity quantify the elements incorporated into the process flow chart. We distinguish between operation elements, which are involved whenever the work part is on the machine and correspond to the circles in the process flow chart, and non-operation elements, which include storages, transportations, delays, and inspections. Letting T_o

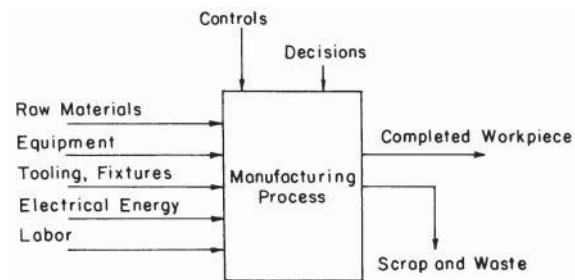


Figure 3 General input–output model of manufacturing process.

represent operation time per machine, T_{no} the nonoperation time associated with each operation, and n_m the number of machines or operations through which each part must be processed, the total time required to process the part through the plant (called the manufacturing lead time, T_l) is

$$T_l = n_m(T_o + T_{no})$$

If there is a batch of p parts,

$$T_l = n_m(pT_o + T_{no})$$

If a setup of duration T_{su} is required for each batch,

$$T_l = n_m(T_{su} + pT_o + T_{no})$$

The total batch time per machine, T_b , is given by

$$T_b = T_{su} + pT_o$$

The average production time T_a per part is therefore

$$T_a = \frac{T_{su} + pT_o}{p}$$

The average production rate for each machine is

$$R_a = 1/T_a$$

As an example, a part requires six operations (machines) through the machine shop. The part is produced in batches of 100. A setup of 2.5 h is needed. Average operation time per machine is 4.0 min. Average nonoperation time is 3.0 h. Thus,

$$n_m = 6 \text{ machines}$$

$$p = 100 \text{ parts}$$

$$T_{su} = 2.5 \text{ h}$$

$$T_o = 4/60 \text{ h}$$

$$T_{no} = 3.0 \text{ h}$$

Therefore, the total manufacturing lead time for this batch of parts is

$$T_l = 6[2.5 + 100(0.06667) + 3.0] = 73.0 \text{ h}$$

If the shop operates on a 40-h week, almost two weeks are needed to complete the order.

3 NUMERICAL CONTROL MANUFACTURING SYSTEMS

3.1 Numerical Control

The most commonly accepted definition of NC is that given by the Electronic Industries Association (EIA): a system in which motions are controlled by the direct insertion of numerical data at some point. The system must automatically interpret at least some portion of these data.

The NC system consists of five basic, interrelated components:

1. Data input devices
2. Machine control unit
3. Machine tool or other controlled equipment
4. Servodrives for each axis of motion
5. Feedback devices for each axis of motion

The major components of a typical NC machine tool system are shown in Fig. 4.

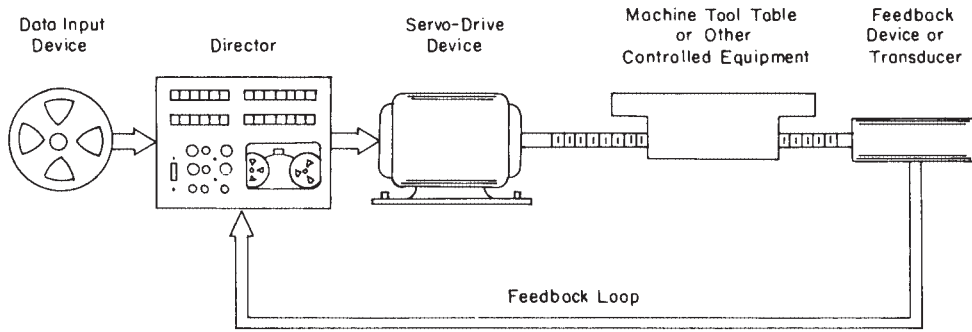


Figure 4 Simplified NC system.

The programmed codes that the machine control unit (MCU) can read may be perforated tape or punched tape, magnetic tape, tabulating cards, or signals directly from computer logic or some computer peripherals, such as disk or drum storage. Direct computer control (DCC) is the most recent development and one that affords the help of a computer in developing a part program.

3.2 Coordinate System

The Cartesian coordinate system is the basic system in NC. The three primary linear motions for an NC machine are given as X , Y , and Z . Letters A , B , and C indicate the three rotational axes, as in Fig. 5.

NC machine tools are commonly classified as being either point to point or continuous path. The simplest form of NC is the point-to-point machine tool used for operations such as drilling, tapping, boring, punching, spot welding, or other operations that can be completed at a fixed coordinate position with respect to the workpiece. The tool does not contact the workpiece until the desired coordinate position has been reached; consequently, the exact path by which this position is reached is not important.

With continuous-path (contouring) NC (CNC) systems, there is contact between the workpiece and the tool as the relative movements are made. Continuous-path NC systems are used primarily for milling and turning operations that can profile and sculpture workpieces. Other NC continuous-path operations include flame cutting, sawing, grinding, and welding and even operations such as the application of adhesives. We should note that continuous-path systems

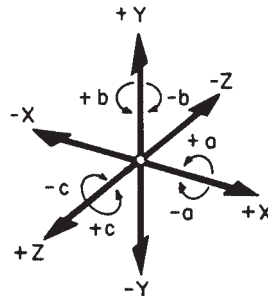


Figure 5 Example of typical axis nomenclature for machine tools.

can be programmed to perform point-to-point operations, although the reverse (while technically possible) is infrequently done.

3.3 Selection of Parts for NC Machining

Parts selection for NC should be based on an economic evaluation, including scheduling and machine availability. Economic considerations affecting NC part selection include alternative methods, tooling, machine loadings, manual versus computer-assisted part programming, and other applicable factors.

Thus, NC should be used only where it is more economical or does the work better or faster or where it is more accurate than other methods. The selection of parts to be assigned to NC has a significant effect on its payoff. The following guidelines, which may be used for parts selection, describe those parts for which NC may be applicable:

1. Parts that require *substantial tooling costs* in relation to the total manufacturing costs by conventional methods
2. Parts that require *long setup times* compared to the machine run time in conventional machining
3. Parts that are machined in *small or variable lots*
4. A *wide diversity of parts* requiring frequent changes of machine setup and a large tooling inventory if conventionally machined
5. Parts that are *produced at intermittent times* because demand for them is cyclic
6. Parts that have *complex configurations* requiring close tolerances and intricate relationships
7. Parts that have *mathematically defined complex contours*
8. Parts that require *repeatability* from part to part and lot to lot
9. *Very expensive* parts where human error would be very costly and increasingly so as the part nears completion
10. *High-priority* parts where lead time and flow time are serious considerations
11. Parts with *anticipated design changes*
12. Parts that involve a *large number of operations* or *machine setups*
13. Parts where *nonuniform cutting conditions* are required
14. Parts that require *100% inspection* or require measuring many checkpoints, resulting in high inspection costs
15. *Family of parts*
16. *Mirror-image parts*
17. *New parts* for which conventional tooling does not already exist
18. Parts that are suitable for *maximum machining* on NC machine tools

3.4 CAD/CAM Part Programming

Computer-aided design consists of using computer software to produce drawings of parts or products. These drawings provide the dimensions and specifications needed by the machinist to produce the part or product. Some well-known CAD software products include *AutoCAD*, *Cadkey*, and *Mastercam*.

Computer-aided manufacturing involves the use of software by NC programmers to create programs to be read by a CNC machine in order to manufacture a desired shape or surface. The end product of this effort is an NC program stored on disk, usually in the form of G codes, that

when loaded into a CNC machine and executed will move a cutting tool along the programmed path to create the desired shape. If the CAM software has the means of creating geometry, as opposed to importing the geometry from a CAD system, it is called *CAD/CAM*. CAD/CAM software, such as Mastercam, is capable of producing instructions for a variety of machines, including lathes, mills, drilling and tapping machines, and wire electrostatic discharge machining (EDM) processes.

3.5 Programming by Scanning and Digitizing

Programming may be done directly from a drawing, model, pattern, or template by digitizing or scanning. An optical reticle or other suitable viewing device connected to an arm is placed over the drawing. Transducers will identify the location and translate it either to a tape puncher or other suitable programming equipment. Digitizing is used in operations such as sheet-metal punching and hole drilling. A scanner enables an operator to program complex free-form shapes by manually moving a tracer over the contour of a model or premachined part. Data obtained through the tracer movements are converted into tape by a minicomputer. Digitizing and scanning units have the capability of editing, modifying, or revising the basic data gathered.

3.6 Adaptive Control

Optimization processes have been developed to improve the operational characteristics of NC machine tool systems. Two distinct methods of optimization are adaptive control and machinability data prediction. Although both techniques have been developed for metal-cutting operations, adaptive control finds application in other technological fields.

The adaptive control (AC) system is an evolutionary outgrowth of NC. AC optimizes an NC process by sensing and logically evaluating variables that are not controlled by position and velocity feedback loops. Essentially, an adaptive control system monitors process variables, such as cutting forces, tool temperatures, or motor torque, and alters the NC commands so that optimal metal removal or safety conditions are maintained.

A typical NC configuration (Fig. 6*a*) monitors position and velocity output of the servosystem using feedback data to compensate for errors between command response. The AC feedback loop (Fig. 6*b*) provides sensory information on other process variables, such as workpiece–tool air gaps, material property variations, wear, cutting depth variations, or tool deflection. This information is determined by techniques such as monitoring forces on the cutting tool, motor torque variations, or tool–workpiece temperatures. The data are processed by an adaptive controller that converts the process information into feedback data to be incorporated into the machine control unit output.

3.7 Machinability Data Prediction

The specification of suitable feeds and speeds is essentially in conventional and NC cutting operations. Machinability data are used to aid in the selection of metal-cutting parameters based on the machining operation, the tool and workpiece material, and one or more production criteria. Techniques used to select machinability data for conventional machines have two important drawbacks in relation to NC applications: Data are generally presented in a tabular form that requires manual interpolation, checkout, and subsequent revisions; and tests on the machine tool are required to find optimum conditions.

Specialized machinability data systems have been developed for NC application to reduce the need for machinability data testing and to decrease expensive NC machining time. Part programming time is also reduced when machinability information is readily available.

A typical process schematic showing the relationship between machinability data and NC process flow is illustrated in Fig. 7.

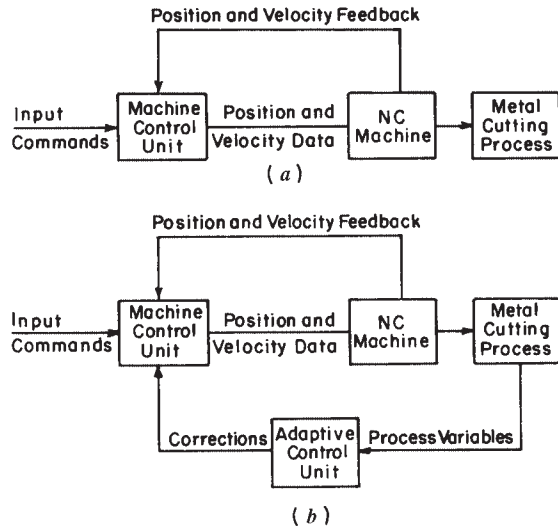


Figure 6 Schematic diagrams for conventional and adaptive NC systems.

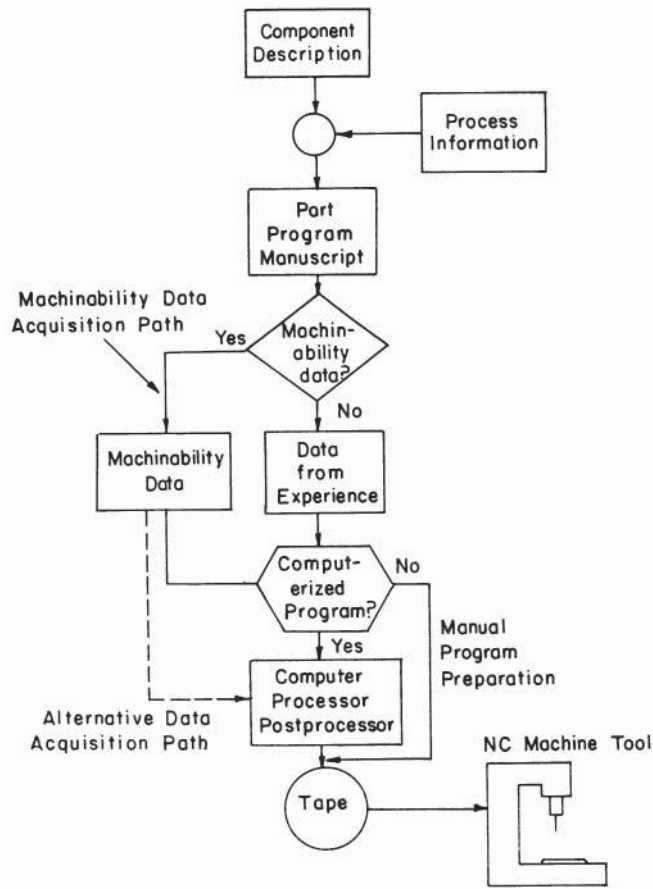


Figure 7 Acquisition of machinability data in NC process flow.

4 INDUSTRIAL ROBOTS

4.1 Definition

As defined by the Robot Institute of America, “a robot is a reprogrammable, multifunctional manipulator designed to handle material, parts, tools, or specialized devices through variable programmed motions for the performance of a variety of tasks.” Robots have the following components:

1. *Manipulator.* The mechanical unit or “arm” that performs the actual work of the robot, consisting of mechanical linkages and joints with actuators to drive the mechanism directly through gears, chains, or ball screws.
2. *Feedback Devices.* Transducers that sense the positions of various linkages or joints and transmit this information to the controller.
3. *Controller.* Computer used to initiate and terminate motion, store data for position and sequence, and interface with the system in which the robot operates.
4. *Power Supply.* Electric, pneumatic, and hydraulic power systems used to provide and regulate the energy needed for the manipulator’s actuators.

4.2 Robot Configurations

Industrial robots have one of three mechanical configurations, as illustrated in Fig. 8. Cylindrical coordinate robots have a work envelope that is composed of a portion of a cylinder. Spherical coordinate robots have a work envelope that is a portion of a sphere. Jointed-arm robots have a work envelope that approximates a portion of a sphere. There are six motions or degrees of freedom in the design of a robot—three arm and body motions and three wrist movements.

Arm and body motions:

1. *Vertical traverse*—an up-and-down motion of the arm
2. *Radial traverse*—an in-and-out motion of the arm
3. *Rotational traverse*—rotation about the vertical axis (right or left swivel of the robot body)

Wrist motions:

4. *Wrist swivel*—rotation of the wrist
5. *Wrist bend*—up-and-down movement of the wrist
6. *Wrist yaw*—right or left swivel of the wrist

The mechanical hand movement, usually opening and closing, is not considered one of the basic degrees of freedom of the robot.

4.3 Robot Control and Programming

Robots can also be classified according to type of control. Point-to-point robot systems are controlled from one programmed point in the robot’s control to the next point. These robots are characterized by high load capacity, large working range, and relative ease of programming. They are suitable for pick-and-place, material handling, and machine loading tasks.

Contouring robots, on the other hand, possess the capacity to follow a closely spaced locus of points that describe a smooth, continuous path. The control of the path requires a large memory to store the locus of points. Continuous-path robots are therefore more expensive than

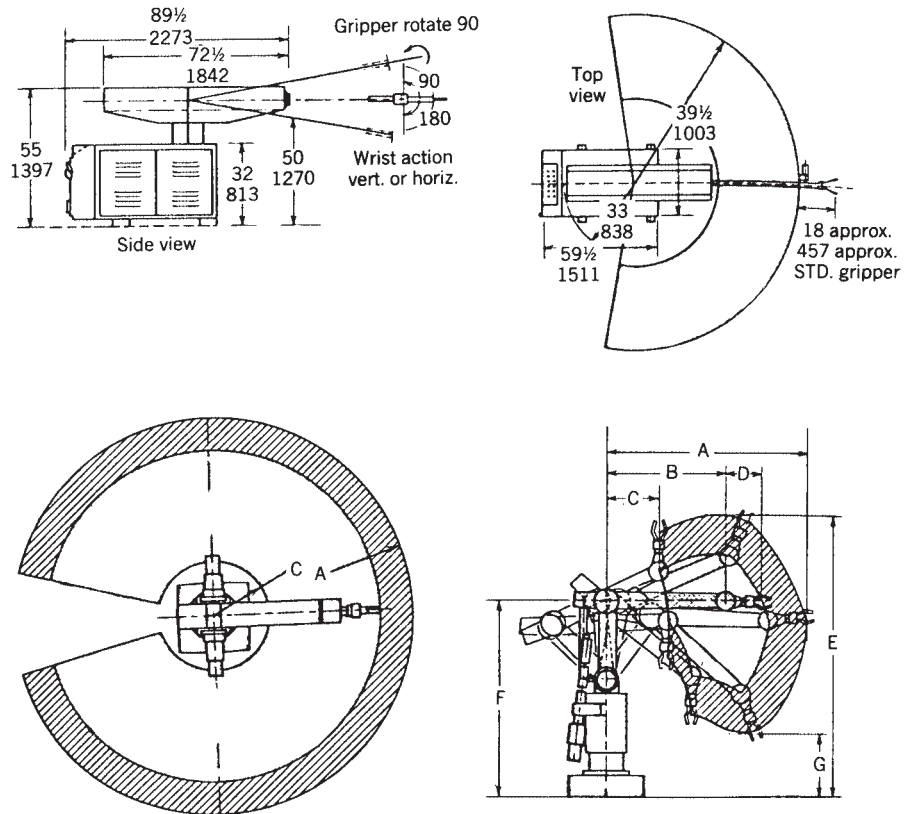


Figure 8 Mechanical configurations of industrial robots.

point-to-point robots, but they can be used in such applications as seam welding, flame cutting, and adhesive beading.

There are three principal systems for programming robots:

1. *Manual Method.* Used in older, simpler robots, the program is set up by fixing stops, setting switches, and so on.
2. *Walk-Through.* The programmer “teaches” the robot by actually moving the hand through a sequence of motions or positions, which are recorded in the memory of the computer.
3. *Lead-Through.* The programmer drives the robot through a sequence of motions or positions using a console or teach pendant. Each move is recorded in the robot’s memory.

4.4 Robot Applications

A current directory of robot applications in manufacturing includes the following:

1. Material handling
2. Machine loading and unloading

3. Die casting
4. Investment casting
5. Forging and heat treating
6. Plastic molding
7. Spray painting and electroplating
8. Welding (spot welding and seam welding)
9. Inspection
10. Assembly

Research and development efforts are under way to provide robots with sensory perception, including voice programming, vision, and “feel.” These capabilities will no doubt greatly expand the inventory of robot applications in manufacturing.

5 COMPUTERS IN MANUFACTURING

Flexible manufacturing systems combined with automatic assembly and product inspection, on the one hand, and integrated CAD/CAM systems, on the other hand, are the basic components of the CIMS. The overall control of such systems is predicated on hierarchical computer control, such as illustrated in Fig. 9.

5.1 Hierarchical Computer Control

The lowest level of the hierarchical computer control structure illustrated in Fig. 9 contains stand-alone computer control systems of manufacturing processes and industrial robots. The computer control of processes includes all types of CNC machine tools, welders, electrochemical machining (ECM), electrical discharge machining (EDM), and laser-cutting machines.

When a set of NC or CNC machine tools is placed under the direct control of a single computer, the resulting system is known as a *direct numerical control* (DNC) system. DNC systems can produce several different categories of parts or products, perhaps unrelated to one another. When several CNC machines and one or more robots are organized into a system for the production of a single part or family of parts, the resulting system is called a *manufacturing cell*. The distinction between DNC systems and a manufacturing cell is that in DNC systems the same computer receives data from and issues instructions to several separate machines, whereas in manufacturing cells the computer coordinates the movements of several machines and robots working in concert. The computer receives “completion of job” signals from the machines and issues instructions to the robot to unload the machines and change their tools. The software includes strategies for handling machine breakdowns, tool wear, and other special situations.

The operation of several manufacturing cells can be coordinated by a central computer in conjunction with an automated material handling system. This is the next level of control in the hierarchical structure and is known as a *flexible manufacturing system* (FMS). The FMS receives incoming workpieces and processes them into finished parts, completely under computer control.

The parts fabricated in the FMS are then routed on a transfer system to automatic assembly stations, where they are assembled into subassemblies or final product. These assembly stations can also incorporate robots for performing assembly operations. The subassemblies and final product may also be tested at automatic inspection stations.

As shown in Fig. 9, FMS, automatic assembly, and automatic inspection are integrated with CAD/CAM systems to minimize production lead time. These four functions are coordinated by

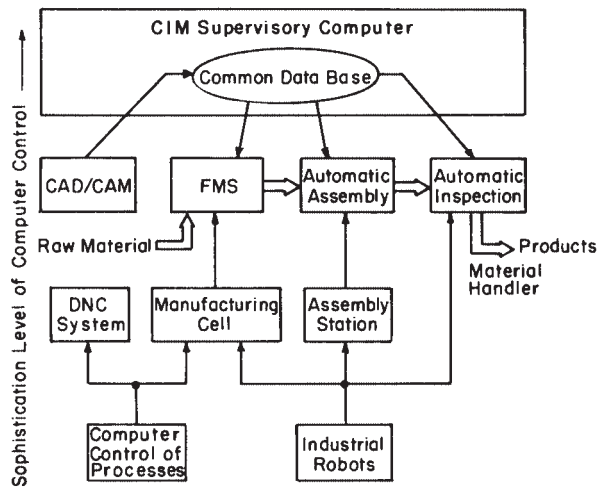


Figure 9 Hierarchical computer control in manufacturing.

means of the highest level of control in the hierarchical structure CIMs. The level of control is often called *supervisory computer control*.

The increase in productivity associated with CIMs will come not from a speedup of machining operations but rather from minimizing the direct labor employed in the plant. Substantial savings will also be realized from reduced inventories, with reductions in the range of 80–90%.

5.2 CNC and DNC Systems

The distinguishing feature of a CNC system is a dedicated computer, usually a microcomputer, associated with a single machine tool, such as a milling machine or a lathe. Programming the machine tools is managed through punched or magnetic tape or directly from a keyboard. DNC is another step beyond CNC, in that a number of CNC machines, ranging from a few to as many as 100, are connected directly to a remote computer. NC programs are downloaded directly to the CNC machine, which then processes a prescribed number of parts.

5.3 Manufacturing Cell

The concept of a manufacturing cell is based on the notion of cellular manufacturing, wherein a group of machines served by one or more robots manufactures one part or one part family. Figure 10 depicts a typical manufacturing cell consisting of a CNC lathe, a CNC milling machine, a CNC drill, open conveyor to bring work parts into the cell, another to remove completed parts from the cell, and a robot to serve all these components. Each manufacturing cell is self-contained and self-regulating. The cell is usually made up of 10 or fewer machines. Those cells that are not completely automated are usually staffed with fewer personnel than machines, with each operator trained to handle several machines or processes.

5.4 Flexible Manufacturing Systems

Flexible manufacturing systems combine many different automation technologies into a single production system. These include NC and CNC machine tools, automatic material handling between machines, computer control over the operation of the material handling system and

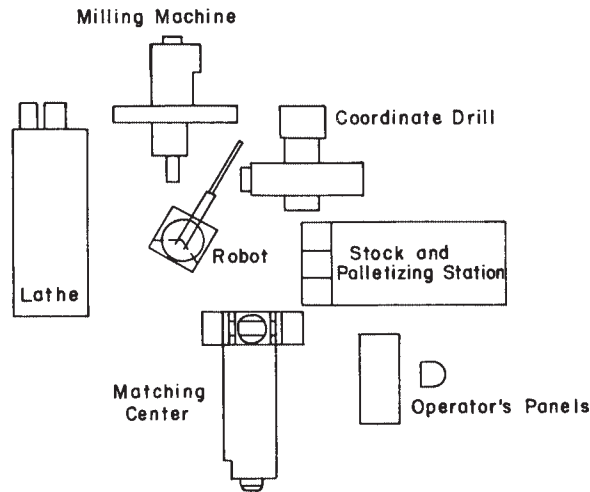


Figure 10 Typical manufacturing cell.

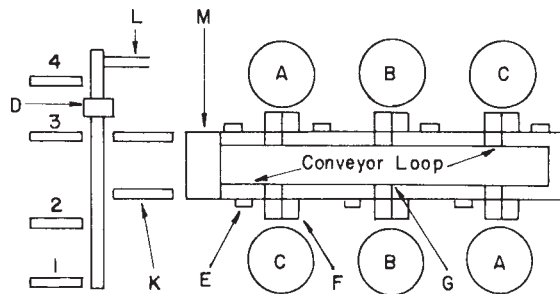


Figure 11 A flexible manufacturing system.

machine tools, and group technology principles. Unlike the manufacturing cell, which is typically dedicated to the production of a single parts family, the FMS is capable of processing a variety of part types simultaneously under NC control at the various workstations.

Human labor is used to perform the following functions to support the operation of the FMS:

- Load raw work parts into the system
- Unload finished work parts from the system
- Change tools and tool settings
- Equipment maintenance and repair

Robots can be used to replace human labor in certain areas of these functions, particularly those involving material or tool handling. Figure 11 illustrates a sample FMS layout.

6 GROUP TECHNOLOGY

Group technology is a manufacturing philosophy in which similar parts are identified and grouped together to take advantage of similarities in design and/or manufacture. Similar parts

are grouped into part families. For example, a factory that produces as many as 10,000 different part numbers can group most of these parts into as few as 50 distinct part families. Since the processing of each family would be similar, the production of part families in dedicated manufacturing cells facilitates workflow. Thus, group technology results in efficiencies in both product design and process design.

6.1 Part Family Formation

The key to gaining efficiency in group-technology-based manufacturing is the formation of part families. A part family is a collection of parts that are similar either due to geometric features such as size and shape or because similar processing steps are required in their manufacture. Parts within a family are different but sufficiently similar in their design attributes (geometric size and shape) and/or manufacturing attributes (the sequence of processing steps required to make the part) to justify their identification as members of the same part family.

The biggest problem in initiating a group-technology-based manufacturing system is that of grouping parts into families. Three methods for accomplishing this grouping are

1. *Visual Inspection.* This method involves looking at the part, a photograph, or a drawing and placing the part in a group with similar parts. It is generally regarded as the most time-consuming and least accurate of the available methods.
2. *Parts Classification and Coding.* This method involves examining the individual design and/or manufacturing attributes of each part, assigning a code number to the part on the basis of these attributes, and grouping similar code numbers into families. This is the most commonly used procedure for forming part families.
3. *Production flow Analysis.* This method makes use of the information contained on the routing sheets describing the sequence of processing steps involved in producing the part, rather than part drawings. Work parts with similar or identical processing sequences are grouped into a part family.

6.2 Parts Classification and Coding

As previously stated, parts classification and coding comprise the most frequently applied method for forming part families. Such a system is useful in both design and manufacture. In particular, parts coding and classification and the resulting coding system provide a basis for interfacing CAD and CAM in CIMs. Parts classification systems fall into one of three categories:

1. Systems based on part design attributes
 - Basic external shape
 - Basic internal shape
 - Length–diameter ratio
 - Material type
 - Part function
 - Major dimensions
 - Minor dimensions
 - Tolerances
 - Surface finish

2. Systems based on part manufacturing attributes

- Primary process
- Minor processes
- Major dimensions
- Length–diameter ratio
- Surface finish
- Machine tool
- Operation sequence
- Production time
- Batch size
- Annual production requirement
- Fixtures needed
- Cutting tools

3. Systems based on a combination of design and manufacturing attributes

The part code consists of a sequence of numerical digits that identify the part's design and manufacturing attributes. There are two basic structures for organizing this sequence of digits:

1. Hierarchical structures in which the interpretation of each succeeding digit depends on the value of the immediately preceding digit
2. Chain structures in which the interpretation of each digit in the sequence is positionwise fixed

The Opitz system is perhaps the best known coding system used in parts classification and coding. The code structure is

12345 6789 ABCD

The first nine digits constitute the basic code that conveys both design and manufacturing data. The first five digits, 12345, are called the *form code* and give the primary design attributes of the part. The next four digits, 6789, constitute the *supplementary code* and indicate some of the manufacturing attributes of the part. The next four digits, ABCD, are called the *secondary code* and are used to indicate the production operations of type and sequence. Figure 12 gives the basic structure for the Opitz coding system. Note that digit 1 establishes two primary categories of parts, rotational and nonrotational, among nine separate part classes.

The MICLASS (Metal Institute Classification System) was developed by the Netherlands Organization for Applied Scientific Research to help automate and standardize a number of design, manufacturing, and management functions. MICLASS codes range from 12 to 30 digits, with the first 12 constituting a universal code that can be applied to any part. The remaining 18 digits can be made specific to any company or industry. The organization of the first 12 digits is as follows:

Digit 1	Main shape
Digits 2 and 3	Shape elements
Digit 4	Position of shape elements
Digits 5 and 6	Main dimensions
Digit 7	Dimension ratio
Digit 8	Auxiliary dimension
Digits 9 and 10	Tolerance codes
Digits 11 and 12	Material codes

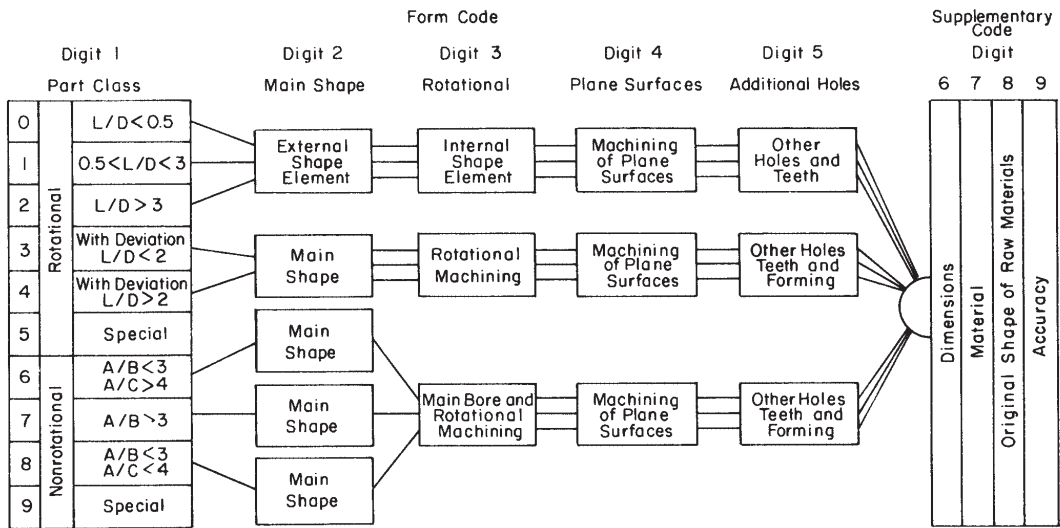


Figure 12 Opitz parts classification and coding system.

MICLASS allows computer-interactive parts coding, in which the user responds to a series of questions asked by the computer. The number of questions asked depends on the complexity of the part and ranges from as few as 7 to more than 30, with an average of about 15.

6.3 Production Flow Analysis

Production flow analysis (PFA) is a method for identifying part families and associated grouping of machine tools. PFA is used to analyze the operations sequence of machine routing for the parts produced in a shop. It groups parts that have similar sequences and routings into a part family. PFA then establishes machine cells for the producing part families. The PFA procedure consists of the following steps:

1. Data collection, gathering part numbers and machine routings for each part produced in the shop
2. Sorting process routings into “packs” according to similarity
3. Constructing a PFA chart, such as depicted in Fig. 13, that shows the process sequence (in terms of machine code numbers) for each pack (denoted by a letter)
4. Analysis of the PFA chart in an attempt to identify similar packs. This is done by rearranging the data on the original PFA chart into a new pattern that groups packs having similar sequences. Figure 14 shows the rearranged PFA chart. The machines grouped together within the blocks in this figure form logical machine cells for producing the resulting part family.

6.4 Types of Machine Cell Designs

The organization of machines into cells, whether based on parts classification and coding or PFA, follows one of three general patterns:

1. Single-machine cell
2. Group machine layout
3. Flow line cell layout

The single-machine pattern can be used for work parts whose attributes allow them to be produced using a single process. For example, a family composed of 40 different machine bolts can be produced on a single turret lathe.

The group machine layout is illustrated in Fig. 13. The cell contains the necessary grouping of machine tools and fixtures for processing all parts in a given family, but material handling

Part No. Machine	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Lathe	x	x		x	x		x	x	x		x	x		x	x		x	x	x	x
Milling Mach. I	x	x	x		x	x	x		x		x		x	x		x				x
Milling Mach. II			x	x				x		x		x	x		x		x	x	x	
Drilling Mach.	x	x	x	x			x	x	x		x	x	x	x		x	x	x		x
Grinding Mach.	x	x	x	x		x			x			x	x		x				x	x

Figure 13 PFA chart.

Part No.																				
Machine	1	2	20	7	11	14	9	5	4	18	12	8	17	15	19	3	13	6	16	10
Lathe	x	x	x	x	x	x	x	x												
Milling Mach. I	x	x	x	x	x	x	x	x												
Drilling Mach.	x	x	x	x	x	x														
Grinding Mach.	x	x	x				x													
Lathe									x	x	x	x	x	x	x					
Milling Mach. II									x	x	x	x	x	x	x					
Drilling Mach.									x	x	x	x	x							
Grinding Mach.									x	x	x			x						
Milling Mach. I																x	x	x	x	
Milling Mach. II																x	x			x
Drilling Mach.																x	x	x	x	x
Grinding Mach.																x	x	x		

Figure 14 Rearranged PFA chart.

between machines is not fixed. The flow line cell design likewise contains all machine tools and fixtures needed to produce a family of parts, but these are arranged in a fixed sequence with conveyors providing the flow of parts through the cell.

6.5 Computer-Aided Process Planning

Computer-aided process planning involves the use of a computer to automatically generate the operation sequence (routing sheet) based on information about the work part. CAPP systems require some form of classification and coding system, together with standard process plans for specific part families. The flow of information in a CAPP system is initiated by having the user enter the part code for the work part to be processed. The CAPP program then searches the part family matrix file to determine if a match exists. If so, the standard machine routing and the standard operation sequence are extracted from the computer file. If no such match exists, the user must then search the file for similar code numbers and manually prepare machine routings and operation sequences for dissimilar segments. Once this process has been completed, the new information becomes part of the master file so that the CAPP system generates an ever-growing data file.

BIBLIOGRAPHY

M. P. Groover, *Automation, Production Systems and Computer-Integrated Manufacturing*, Prentice-Hall, Upper Saddle River, NJ, 2001.

M. P. Groover, M. Weiss, R. N. Nagel, and N. G. Odrey, *Industrial Robotics: Technology, Programming, and Applications*, McGraw-Hill, New York, 1986.

G. Boothroyd, P. Dewhurst, and W. Knight, *Product Design for Manufacture and Assembly*, Marcel Dekker, New York, 1994.

- J. A. Buzacott and D. D. Yao, "Flexible Manufacturing Systems: A Review of Analytical Models," *Manag. Sci.*, **32**, 890–895, 1986.
- W. M. Chow, *Assembly Line Design*, Marcel Dekker, New York, 1990.
- C. Moodie, R. Uzsoy, and Y. Yih, *Manufacturing Cells: A Systems Engineering View*, Taylor & Francis, London, 1995.
- H. Opitz and H. P. Wiendohl, "Group Technology and Manufacturing Systems for Medium Quantity Production," *Int. J. Production Res.*, **9**, 181–203, 1971.

CHAPTER 12

TRIZ

James E. McMunigal

MCM Associates

Long Beach, California

Steven Ungvari

Strategic Product Innovations, Inc.

Columbus, Ohio

Michael Slocum

Breakthrough Management Group

Longmont, Colorado

Ruth E. McMunigal

MCM Associates

Long Beach, California

1	WHAT IS TRIZ?	362			
2	ORIGINS OF TRIZ	362			
2.1	Altshuller's First Discovery	362			
2.2	Altshuller's Second Discovery	362			
2.3	Altshuller's Third Discovery	363			
2.4	Altshuller's Levels of Inventiveness	363			
3	BASIC FOUNDATIONAL PRINCIPLES	364			
3.1	Ideality	364			
3.2	Contradictions	365			
3.3	Technical Contradictions	366			
3.4	Physical Contradictions	366			
3.5	Maximal Use of Resources	367			
4	A SCIENTIFIC APPROACH	367			
4.1	How TRIZ Works	368			
4.2	Five Requirements for a Solution to Be Inventive	370			
5	CLASSICAL AND MODERN TRIZ TOOLS	370			
5.1	Contradiction Matrix	371			
5.2	Physical Contradictions	371			
5.3	Formulating and Solving Physical Contradictions				373
5.4	An Example				373
5.5	Laws of Systems Evolution				374
5.6	Analytical Tools				375
5.7	Su-Field				376
6	PROBLEMS WITHOUT CONTRADICTIONS				376
7	RULES FOR THE INVENTOR: SU-FIELD SYNTHESIS				378
8	CLASS 4: MEASUREMENT AND DETECTION STANDARDS				379
9	ALGORITHM FOR INVENTIVE PROBLEM SOLVING				383
9.1	Steps in ARIZ				384
10	CAVEAT				388
11	CONCLUSION				388
	BIBLIOGRAPHY				388

1 WHAT IS TRIZ?

TRIZ is the acronym for the Russian words *Teoriya Resheniya Izobretatelskikh Zadatch* (theory of the solution of inventive problems). TRIZ's development, evolution, and refinement cover over 50 years of rigorous, empirically based analysis.

The creativity and innovation mentioned within the context of science are rare. Typically, creativity and innovation are considered spontaneous phenomena occurring in a capricious and unpredictable way. Individuals such as Michelangelo, Leonardo da Vinci, and Thomas Edison appear to have possessed innate, natural ability for creative thought and inventiveness. What characteristics enabled them, or anyone, to perform as a highly creative thinker?

The term *theory to the solution of inventive problems* implies there is an innovation and/or creative thought process (supported by an underlying construct and architecture) that can be deployed on an as-needed basis. The implications of such a theory, if true, are enormous, suggesting that technicians can elevate their creative thinking abilities by orders of magnitude when the need arises.

2 ORIGINS OF TRIZ

The catalyst for TRIZ was a Russian named Genrich Altshuller (1926–1998). His interest in inventions began at an early age, patenting a device for generating oxygen from hydrogen peroxide by age 14. Altshuller's fascination with inventions and innovation continued through Stalin's regime and World War II. After the war, he was assigned as a patent examiner for the Department of the Navy. He found himself helping would-be inventors solve various problems with their inventions. Over time, Altshuller became fascinated with the study of inventions and understanding how their inventors' minds worked. His initial attempts were psychologically based; however, these probes provided little if any insight on how creativity could be "engineered."

Altshuller turned his attention to studying inventions and reverse engineering them to understand the essential engineering problem being solved and the elegance of the solution as described in the patent application. Patent applications, called author certificates (ACs) in the former Soviet Union, were concise documents of three to four pages. The AC consisted of a descriptive title of the invention, a schematic of the new invention, a rendering of the current design, the purpose of the invention, and a description of the solution.

2.1 Altshuller's First Discovery

The brevity of the ACs facilitated analysis, cataloguing, and mapping of solutions to the problems. As the number of inventors applying for an AC increased, Altshuller uncovered similar patterns of solutions for similar problems. He developed a scientific, standardized approach to a problem and incorporated a latent knowledge base as an integral element of the solution process when he recognized that similar technological problems gave rise to similar patents. This phenomenon was repeated in widely disparate engineering disciplines, in various geographical areas, during different time frames.

Altshuller postulated the possibility of creating a mechanism for describing "types" of problems and mapping them to types of solutions. This led to a mechanism naming the 39 typical engineering parameters, the contradiction matrix, and 40 inventive principles.

2.2 Altshuller's Second Discovery

As Altshuller assembled chronological technology maps, he uncovered regularity in the evolution of engineered systems. He described these time-based phenomena as "laws" and called

them the *eight laws of engineered systems evolution*. The term laws does not imply that they conform to a strict scientific construction as one would describe in the field of physics or chemistry. The laws, though general in nature, are recognizable and predictable and provide a road map to future derivatives. Today, these eight laws have been expanded into more than 400 “sublines” of evolution and are useful in technology development, product planning, and the establishment of defensible patent fences.

2.3 Altshuller’s Third Discovery

The third truism that emerged was the realization that inventions are vastly different in their degree of “inventiveness.” Indeed, many of the patents he studied were filed to describe a system and provide some degree of protection. These patents were useless to Altshuller’s determination for discovering the secret of how an inventor reaches the highest order. To differentiate inventiveness, he devised a scale of 1–5 for categorizing the elegance of the solution. See Fig. 1.

Only levels 3 and 4 solutions are deemed “inventive.” Within the body of TRIZ knowledge, “inventive” states that the solution was one that did not compromise conflicting requirements. For example, strength versus weight is an example of conflicting parameters. To increase strength, the engineer will typically make something thicker or heavier. An inventive solution would increase strength with no additional weight or even a reduction in weight.

2.4 Altshuller’s Levels of Inventiveness

Level 1: Parametric Solution

A solution utilizing well-known methods and parameters within an engineering field of specialty is the lowest level solution and is not an inventive solution. For example, a problem with roads and bridges icing over can be solved with the application of salt or sand.

Level	Nature of solution	Number of trials to find the solution	Origin of the solution	% of patents at this level
1	Parametric	None to few	The designer’s field of specialty	32%
2	Significant improvement in paradigm	10–50	Within a branch of technology	45%
3	Inventive solution in paradigm	Hundreds	Several branches of technology	18%
4	Inventive solution out of paradigm	Thousands to tens of thousands	From science—physical/chemical effects	4%
5	True discovery	Millions	Beyond contemporary science	1%

Figure 1 Altshuller’s levels of inventiveness.

Level 2: Significant Improvement in the Technology Paradigm

A significant improvement in the system utilizing known methods possible from several engineering disciplines is a level 2 solution. It is a significant improvement over the previous system, but it is not inventive.

A level 2 solution of the icing problem would be required if conventional means were prohibited. This type of solution demands choosing between several variants that leave the original system essentially intact. The roadways or bridges, for example, could be formulated or coated with an exothermic substance that is triggered at a certain temperature.

Level 3: Invention within the Paradigm

The elimination of conflicting requirements within a system utilizing technologies and methods within the current paradigm is a level 3 solution. It is deemed to be inventive because it eliminates the conflicting parameters in such a way that both requirements are satisfied simultaneously.

A level 3 solution to the conflicting requirements of strength versus weight has been solved in aircraft by the use of honeycomb structures and composites.

Level 4: Invention outside the Paradigm

Creation of a new generation of a system with a solution derived not in technology but in science is a level 4 solution. It integrates several branches of science. The invention of the radio, the integrated circuit, and the transistor are examples of level 4 solutions.

Level 5: True Discovery

A level 5 discovery is one that is beyond the bounds of contemporary science. It will often spawn entire new industries or allow for the accomplishment of tasks in radically new ways. Laser and the Internet are examples of level 5 inventions.

3 BASIC FOUNDATIONAL PRINCIPLES

Altshuller's *three discoveries* provide the construct for the formation of the foundational underpinnings upon which all TRIZ theory, practices, and tools are built. The three building blocks of TRIZ are *ideality*, *contradictions*, and the *maximal use of resources*.

3.1 Ideality

The notion of ideality is a simple concept. Ideality states that, in the course of time, systems move toward a state of increased ideality. Ideality is the ratio of useful functions F_U to harmful functions F_H :

$$\text{Ideality} = I = \frac{\sum F_U}{\sum F_H}$$

Useful functions embody all of the desired attributes, functions, and outputs from the system. From an engineering point of view, it is what is termed design intent. Harmful functions, on the other hand, include the expenses or fees that are associated with the system, the space it occupies, the resources it consumes, cost to manufacture, cost to transport, cost to maintain, etc.

Extrapolating the concept to its theoretical limit, one arrives at a situation where a system's output consists solely of useful functions with the complete absence of any harmful consequences. Altshuller called this state the ideal final result (IFR). The IFR is not calculated; it is

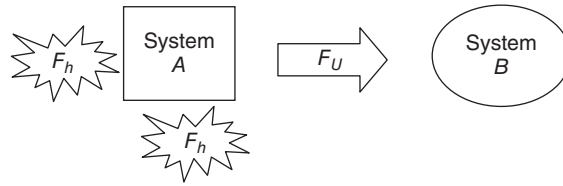


Figure 2 System A and system B interaction with useful output and harmful effects.

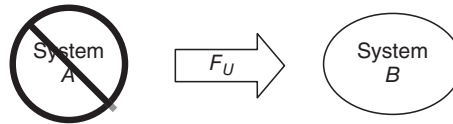


Figure 3 A's effect of A when there is no system A yet system A function is carried out.

a tool to define the ideal end state. Once the end state is defined, the question as to why it is difficult to attain it flushes out the real (contradictory) problems that must be overcome.

One might argue that it is absurd to think of solving problems from the theoretical notion of the IFR instead of explicitly defining the current dimensions of the problem. However, it is precisely this point of view that opens up innovative vistas by reducing prejudice, bias, and psychological inertia (PI).

Psychological inertia is analogous to thinking only within one's paradigm. An engineer competent in mechanics, for example, is unlikely to search for a solution in chemistry because it is outside his or her paradigm.

Problems with long duration yield an especially target-rich environment for TRIZ. Those intelligent folks who own the problem tend to work in their technical domain and the solution space often resides elsewhere. Some examples where discipline lines were successfully crossed are mechanical to microelectronic and composite lay-up to injection mold.

The notion of ideality postulates that a system, any system, is not a goal in itself. It is only a goal or design intent of any system—the useful function(s) that the system provides. Taken to its extreme, the most ideal system is one that does not exist but nevertheless one that produces its intended useful function(s). See Figs. 2 and 3.

In Fig. 3 the system has not reached a state of ideality because the useful interaction between A and B is accompanied by some type of unwanted (harmful) function. An ideal system A, on the other hand, is one that does not exist even when its design intent is fully accomplished.

In the abstract, this notion might seem fantastical or even absurd. There is, however, a very subtle yet very powerful heuristic embodied in ideality. First, ideality creates a mind set for finding a noncompromising solution. Second, it is effective in delineating all of the technological hurdles that need to be overcome in order to invent the best solution possible. Third, it forces the problem solver to find alternative means or resources in order to provide the intended useful function. The latter outcome is similar to an organization reassigning key functions to the individuals that are retained after a reduction in force.

3.2 Contradictions

The second foundation principle is the full recognition that systems are inherently rife with various conflicts. Within TRIZ these conflicts are called contradictions. In TRIZ an “inventive” problem is one that contains one or more contradictions. Typically, when one is faced with a contradictory set of requirements, the easy resolution is to find a compromising solution.

This type of solution, while it may be expedient, is not an inventive solution. If we return to the example of weight versus strength, an inventive solution satisfies both requirements. Another example would be speed versus precision. A TRIZ level 3 solution would satisfy both requirements utilizing available “in-paradigm” methods, while a level 4 solution would incorporate technologies outside of the current paradigm. In both cases, however, speed and precision would be achieved at a quality level demanded by the contextual parameters of the situation. In TRIZ, two distinct types of contradictions are delineated: technical contradictions and physical contradictions. (Methods for solving technical contradictions will be discussed later in the chapter.)

3.3 Technical Contradictions

A technical contradiction is a situation where two identifiable parameters are in conflict: When one parameter is improved, the other is made worse. The two previously mentioned, weight versus strength and speed versus precision, are examples. See Fig. 4.

3.4 Physical Contradictions

A physical contradiction is a situation where a single parameter needs to be in opposite physical states; e.g., it needs to be thin and thick, hot and cold, etc., at the same time. A physical contradiction is the controlling element or parameter linking the parameters of the technical contradiction. Figure 5 shows the pulley (C) upon which parameters A and B rotate as the physical contradiction.

The physical contradiction lies at the heart of an inventive problem; it is the ultimate contradiction. When the physical contradiction has been found, the process of generating an inventive

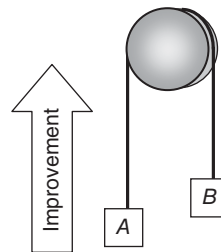


Figure 4 Relationship of parameters A and B; as one improves, the other worsens.

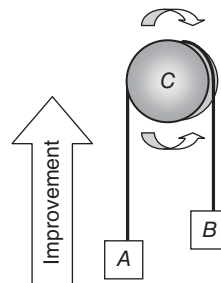


Figure 5 Physical contradictions example. For A and B to improve, C must rotate clockwise and counter-clockwise simultaneously.

Table 1 Types of Resources

Substance	—any material contained in the system or its environment, manufactured products, or wastes
Energy	—any kind of energy existing in the system
Space	—any space available in the system and its environment
Time	—time intervals before start, after finish, and between technological cycles, unused or partially used
Functional	—possibilities of the system or its environment to carry out additional functions
Own	—unused specific features and properties, characteristics of a particular system, such as special physical, chemical, or geometric properties; for example, resonance frequencies, magnetosusceptibility, radioactivity, and transparency at certain frequencies
System	—new useful functions or properties of the system that can be achieved from modification of connections between the subsystems or a new way of combining systems
Organizational	—existing but incompletely used structures or structures that can be easily built in the system, arrangement or orientation of elements, or communication between them
Differential	—differences in magnitude of parameters that can be used to create flux that carry out useful functions; for example, speed difference for steam next to a pipe wall versus in the middle, temperature variances, voltage drop across resistance, height variance
Changes	—new properties or features of the system (often unexpected), appearing after changes have been introduced
Harmful	—wastes of the system (or other systems) which become harmless after use

solution has been greatly simplified. It stands to reason that when a physical contradiction is made to behave in two opposite states simultaneously, the technical contradiction is eliminated. For example, if by some means pulley *C* could rotate in opposite directions at the same time, both *A* and *B* would increase, hence eliminating the technical contradiction.

3.5 Maximal Use of Resources

The third foundation principle of TRIZ is the maximal utilization of any available resources before introducing a new component or complication into the system.

Resources are defined as any substance, space, or energy that is present in the system, its surroundings, or the environment. The identification and utilization of resources increase the operating efficiency of the system, thereby improving its ideality. In the former Soviet Union, where money was scarce, necessity proved to be the mother of invention. In the West, on the other hand, system problems were often engineered out by “throwing money and complexity” at a system. The utilization of resources as an X agent to solve the problem was and still is not widely practiced.

A practiced TRIZ problem solver will marshal any in-system or environmental resources to assist in solving the problem. It is only when all resources have been exhausted or it is impractical to utilize one that the consideration for additional design elements comes into play. The mantra of a TRIZ problem solver is, “*Never solve a problem by making the system more complex.*” More on this when the algorithm for problem solving (ARIZ) is discussed. Table 1 lists the types of resources used in TRIZ.

4 A SCIENTIFIC APPROACH

TRIZ is comprised of a comprehensive set of analytical and knowledge-based tools that were heretofore buried at a subconscious level in the minds of creative inventors. If asked to explain specifically how they invent, most people are unable to provide a repeatable formula. Through his work, Altshuller has codified the amorphous process of invention. His contribution to society is that he made the process of inventive thinking explicit. He has made it possible for anyone

with a reasonable amount of intelligence to become an inventor. TRIZ makes it possible for people of average intelligence to access a large body of inventive knowledge and through analogical analysis formulate inventive “out-of-the-box” solutions.

4.1 How TRIZ Works

The general scheme in TRIZ is solution by abstraction. A specific problem is described into a more abstract form. The abstracted form of the problem has a counterpart solution at the level of abstraction. The connection between the problem and the solution is found through the use of various TRIZ tools. Once the solution analog is arrived at, the process is reversed, producing a specific solution. Figure 6 illustrates the process of solution by abstraction and Fig. 7 applies the process to an algebraic problem.

Assume that we were given the task of solving the problem found in the equation $3x^2 + 5x + 2 = 0$. Without a specific process, we would be reduced to the inefficient process of trial and error. An even more absurd method would be to try to arrive at the answer by brainstorming. Yet, brainstorming is often applied to problems that are much more complex than that shown above. This is what makes TRIZ so compelling—it provides a roadmap to highly creative and innovative solutions to seemingly impossible problems.

Figure 7 provides the general schema for how TRIZ works. The fundamental idea in TRIZ is to reformulate a problem into a more general (abstract) problem and then find an equivalent “solved” problem. These analogs, in theory, define the solution space that is occupied by one or several noncompromising alternative solutions.

The advantage of increasing the level of abstraction is that the solution space is expanded. Solving the equation illustrated in Fig. 7 is relatively simple, assuming knowledge of algebra. The correctness of the solution is also easier to verify because the solution space is very small, e.g., there is only one right answer! Inventive problems pose a greater challenge than the one shown above because the solution space is very large. Figure 8 illustrates this truism.

Figure 8 shows what happens when solving “inventive” versus noninventive problems. An inventive problem is often confused with problems of design, engineering, or a technological nature. For example, in constructing a bridge, the type of bridge to be built is largely an issue

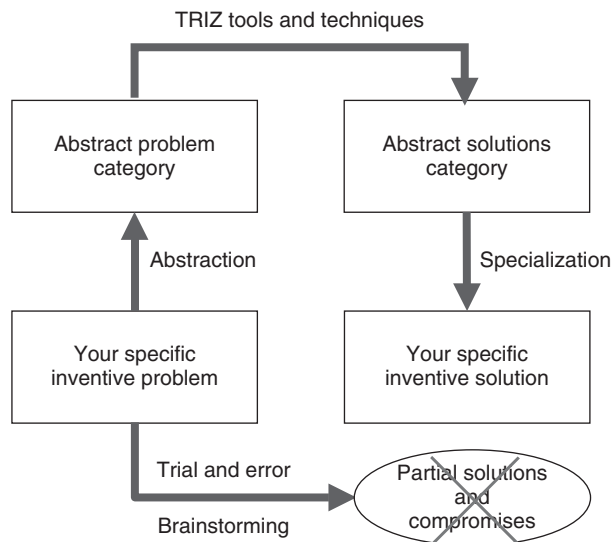


Figure 6 Solution by abstraction using TRIZ.

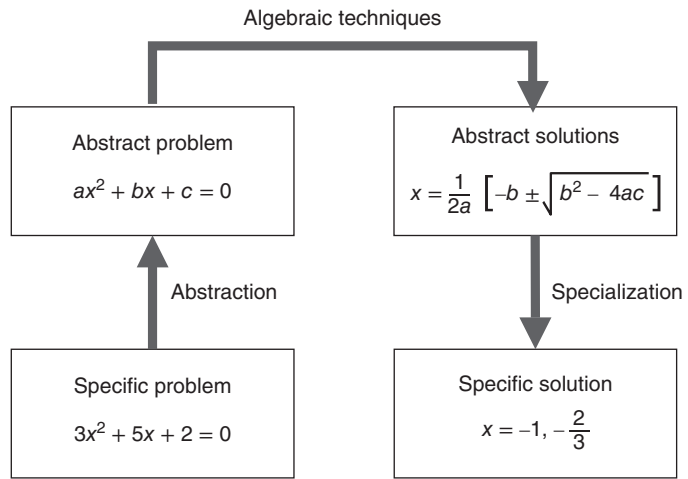


Figure 7 Solution by abstraction using algebraic techniques.

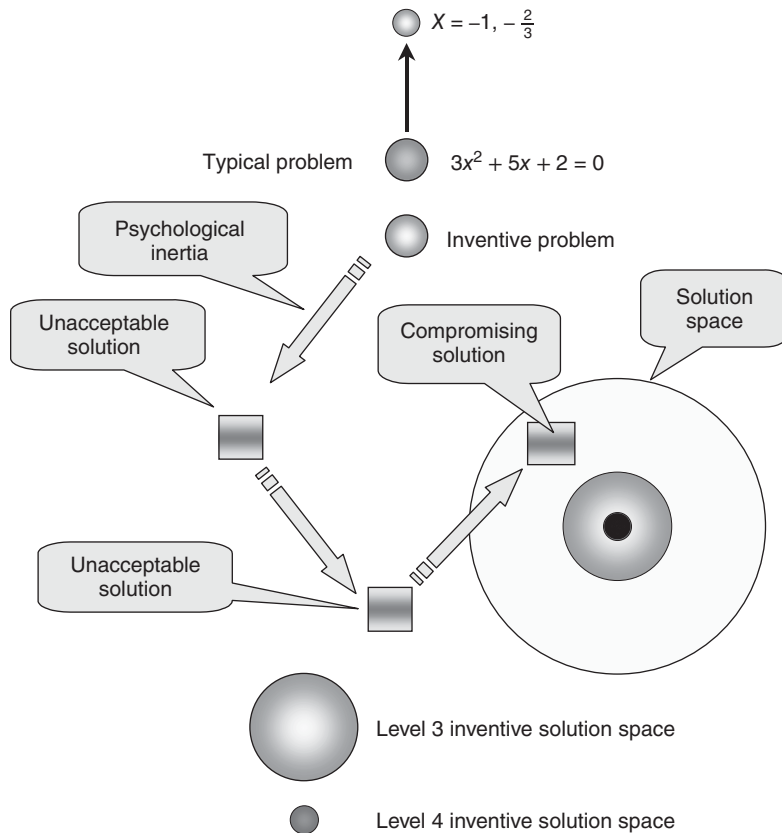


Figure 8 Solution space for inventive versus other problems.

Table 2 Requirements of Inventive Solutions

Solution fully resolves the contradictory requirements.
Solution preserves all advantages of the current system.
Solution eliminates the disadvantages of the current system.
Solution does not introduce any new disadvantages.
Solution does not make system more complex.

related to design. A cantilever bridge provides known design advantages over a suspension bridge in specific contexts and vice versa. This is an example of a noninventive design problem. The calculations of load and stress the bridge will have to withstand are an engineering problem. Coordinating the construction and assuring that materials meet specifications and the job is on time and within budget is a technical problem. While any of these problems are not insignificant by themselves, they are not inventive within the context of TRIZ because they are solvable by using known methods, formulas, schedules, etc. Furthermore, the path to the correct solution is defined and direct, and since the solution space is very small, verification of the answer is straightforward. This is not the case with inventive problems.

An inventive problem in the context of building a bridge would be to make the bridge lighter and stronger, larger and less expensive, longer and more stable, etc. These problems are inventive because they have to overcome one or more contradictions. Therefore, to reiterate, a problem is an inventive one when one or several contradictions must be overcome in the solution and a compromise solution is not acceptable.

There are several distinguishing characteristics of an inventive versus typical problem, as shown in Fig. 8. The entire solution space can be quite large, containing both noninventive and inventive solutions. The two inner concentric circles represent level 3 and level 4 inventive solutions, while the larger outer circle represents an area of noninventive solutions. Just as it is harder to hit the bull's-eye when shooting an arrow, so it is with hitting on an inventive solution. Why is this so?

The initial factor oftentimes driving one off the mark, PI, defined previously, presupposes a solution path as defined by one's individual paradigms. The route to a solution is often one of trial and error and strewn with several unacceptable solutions arrived at along the vector of one's PI. In a sense, the process of defining the current problem and then driving to a solution can be considered a "push" method for finding a solution. TRIZ is different because one of the initial steps of the TRIZ process is to define the ideal state, i.e., the solution space found in level 3 or 4 solutions. The articulation of the ideal solution acts to orient the problem solver and "pulls" him or her in that direction. Furthermore, TRIZ guides one to the ideal solution through the process of abstraction and finding analogs, as discussed previously. These two fundamental elements of TRIZ serve as a powerful magnet to draw one to an inventive solution and provide an example of how this has been accomplished by a previous inventor.

4.2 Five Requirements for a Solution to Be Inventive

Within the context of TRIZ, before a proposed solution is labeled inventive, it must meet all of the stringent requirements outlined in Table 2.

5 CLASSICAL AND MODERN TRIZ TOOLS

In the course of his analytical work, Altshuller amassed a vast body of knowledge and invented analytical methods on how to access that body of knowledge. The subsequent evolution of TRIZ followed logical parallel paths. The creation of a body of "inventive" knowledge gave rise to

various analytical tools making it easier to catalog and create more inventive knowledge that, in turn, spawned more sophisticated tools, and so on. The end result after more than 50 years of work is a complete set of sophisticated tools and an immense knowledge base of inventive ideas, methods, and solutions that can be mobilized to attack any inventive problem. To date, these tools have been used to solve problems related to product design and development, quality, manufacturing, cost reduction, production, warranty, and prevention of product failures, to name just a few applications.

The tools of TRIZ are subdivided into two major categories. The first division is by the nature of the tool, e.g., analytical versus knowledge base. The second differentiation is chronological: classical TRIZ versus I-TRIZ. The classical TRIZ tools span those derived from 1946 to 1985 with Altshuller as the primary inventive force. A protégée of Altshuller, Boris Zlotin of The Kishnev School (of TRIZ), continued developing the methodology, which for purposes of differentiation is called I-TRIZ.

5.1 Contradiction Matrix

The first of the classical TRIZ tools invented by Altshuller is the contradiction matrix. The objective of the matrix is to direct the problem-solving process to incorporate an idea that has been utilized before to solve an analogous “inventive” problem. The contradiction matrix accomplishes this by asking two simple questions: Which element of the system is in need of improvement? If improved, which element of the system is deteriorated? This is a technical contradiction. A portion of the 39×39 matrix is shown in Fig. 9.

The matrix is constructed by juxtaposing 39 engineering parameters along the vertical and horizontal axis. At the intersections Altshuller filled in from one to four numerical values hinting at ways to solve the problem. The numerical values identified one of the 40 inventive principles that were culled from the knowledge base as ways in which an analog to the specific problem had been solved previously. The 39 engineering parameters are general in nature and act as surrogates for the specific “real” parameters in conflict. The inventive principles are broad and nonspecific as the exact way in which they should be applied. In Fig. 9, the problem is trying to improve “convenience of use,” but when this is attempted, it results in “waste of energy.” The matrix suggests that when this type of problem is encountered, principles 2, 9, and 13 have been utilized to resolve the contradiction. Table 3 provides details on these three principles.

The process for using the contradiction matrix follows the general schema shown in Fig. 7. The steps are as follows:

1. Describe the problem.
2. Select the parameter most closely aligned with one of the 39 engineering parameters from the feature-to-improve column.
3. State your proposed solution.
4. Select which feature will be deteriorated.
5. Note the inventive principle(s) at the intersection.
6. Apply the inventive principle(s).

5.2 Physical Contradictions

A physical contradiction (PC) is the controlling element in the system that links the two conflicting parameters in the technical contradiction; see Fig. 4. The PC expresses the most extreme form of contradictory requirements because the conflict must be resolved solely within a single entity. As Fig. 4 shows, the PC (pulley) is at the root of the inventive problem. If it were possible to make the pulley turn in opposite directions simultaneously, the technical contradiction would

Deteriorated feature / Feature to improve		1	2	3	●	22
		Weight of a moving object	Weight of nonmoving object	Length of a moving object	●	Waste of energy
1	Weight of a moving object			15,8 29,34	●	6, 2 34, 19
2	Weight of a nonmoving object				●	18, 19 28, 15
3	Length of a moving object	8,15 29,34			●	7, 2 35,39
4	Length of a nonmoving object		35,28 40,29		●	6, 28
5	Area of a moving object	2,17 29,4		14,15 18,4	●	15, 17 30,26
6	Area of a nonmoving object		30,2 14,18		●	17, 7 30
7	Volume of moving object	2,26 29,40		1,7 35,4	●	7, 15 13,16
●	● ● ● ● ● ● ●	●	●	●	●	●
33	Convenience of use	25,2 13,15	6,13 1,25	1,17 13,12	●	2,19 13
34	Repairability	2,27 35,11	2,27 35,11	1,28 10,25	●	15, 1 32,19
35	Adaptability	1,6 15,8	19,15 29,16	35,1 29,2	●	18, 15 1
36	Complexity of device	26,30 34,36	2,36 35,39	1,19 26,24	●	10,35 13,2
37	Complexity of control	27,26 28,13	6,13 28,1	16,17 26,24	●	35,3 15,19
38	Level of automation	28,26 18,35	28,26 35,10	14,13 17,28	●	23,28
39	Productivity	35,26 24,37	28,27 15,3	18,4 28,38	●	28,10 29,35

Figure 9 Contradiction matrix.

Table 3 Three of the Forty Inventive Principles

3.	Local quality
a.	Change an object's structure from uniform (homogeneous) to nonuniform (heterogeneous) or change the external environment (or external influence) from uniform to nonuniform
b.	Have different parts of the object carry out different functions
c.	Place each part of the object under conditions most favorable for its operation
9.	Preliminary antiaction
a.	If it is necessary to perform some action with both harmful and useful effects, consider a counteraction in advance that will negate the harmful effects
b.	Create stresses in an object that will counter known undesirable forces later on
13.	The other way around
a.	Instead of an action dictated by the specifications of the problem, implement an opposite action
b.	Make a moving part of the object or the outside environment immovable and the nonmoving part movable
c.	Turn the object upside down and inside out; freeze it instead of boiling it

disappear. From a TRIZ standpoint, solving an inventive problem by satisfying the conflicting requirements of the PC results in elegant solutions with a greater degree of inventiveness.

5.3 Formulating and Solving Physical Contradictions

A PC is formulated according to the logic: To perform function F_1 , the object must exhibit property P , but to perform function F_2 , it must exhibit property $-P$. The solution to PCs is accomplished by incorporating principles of separation. There are five separation principles that can be used to resolve a PC. See Table 4.

5.4 An Example

The principle of separation in time can be explained by a well-known illustration used by Altshuller. Assume that one is driving concrete piles for buildings into very hard ground. To facilitate ease of driving the piles, the tip profile should be sharp. Once in place, the pile should be stable, which means the profile should be blunt. In other words, the pile should be sharp and blunt—a PC. How can this be? The problem is solved by imbedding an explosive into the sharp end of the pile and, when it is in place, destroying the sharp profile by setting off the explosive. The tip profile is sharp (P) during time T_1 (driving into the ground) and it is blunt ($-P$) during time T_2 (in place).

Table 4 Separation Principles

-
1. Separation in time
 2. Separation in space
 3. Separation between the system and its components
 4. Separation upon condition
 5. Coexistence of contradictory properties
-

5.5 Laws of Systems Evolution

The notion of predicting future technological patterns and derivatives has been recognized as a means of creating competitive leverage. Techniques such as technology forecasting, morphological analysis, trend extrapolation, and the Delphi process have been utilized since World War II. All of these techniques are based on statistical probability modeling. In TRIZ, future derivatives are based on predetermined patterns of evolution that have been around since the invention of the wheel. Past evolutionary trends provide an “evolutionary crystal ball” for understanding how current technologies will morph over time. Altshuller termed these phenomena “laws of evolution.”

These laws represent a stable and repeatable pattern of interactions between the system and its environment. These patterns occur because systems are subject to various cycles of improvement. When a new technological system emerges, it typically provides the minimum degree of functionality required to satisfy the inventor’s intent. For example, the first powered flight by the Wright brothers occurred on December 17, 1903. The *Flyer*, with Orville Wright as the pilot, flew to a height of 10 ft and landed heavily after 12 s in the air. Today, jets are capable of flying at heights over 60,000 ft over thousands of miles at several times the speed of sound. What happened with airplanes has been repeated in other types of engineered systems.

The way in which systems evolve can be shown on life-cycle, or S, curves. Figure 10 shows the evolutionary picture.

From the time a system emerges to point *a*, its development is slow as it is unproven. At point *a*, the dominant design paradigm appears and the system is poised for commercialization. From point *a* to point *b* the system experiences rapid improvement as commercialization and market pressures force cycles of continuous improvements. From point *b* to point *c* the rate of improvement slows as the technology matures. As the system passes point *b*, the next system (*B*) is itself emerging. The abandonment of the original system in favor of the new one is governed by how much greater potential it possesses in comparison to the unrealized improvements remaining in system *A*.

Being a keen observer of inventive phenomena, Altshuller, through his analysis, uncovered eight describable, chronologically sequenced events. He called these events the laws of systems evolution. See Table 5.

Within these eight major laws, Altshuller and his students have found numerous “sublines” of evolution. Given the detail that is now captured in the evolutionary knowledge base, it is possible through the analysis of patents to fix where the technological system is positioned on its life-cycle curve.

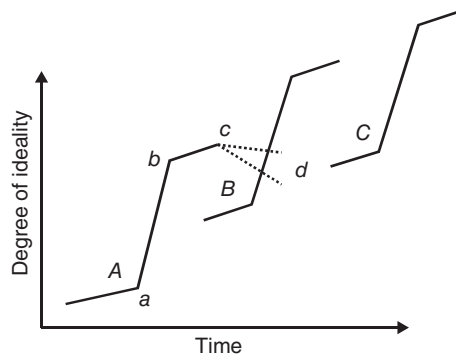


Figure 10 Life-cycle (S) curves, evolutionary picture.

Table 5 Patterns of Technological Systems Evolution

-
1. Stages of evolution
 2. Evolution toward increased ideality
 3. Nonuniform development of systems elements
 4. Evolution toward increased dynamism and controllability
 5. Increased complexity, then simplification
 6. Evolution with matching and mismatching components
 7. Evolution toward microlevel and increased use of fields
 8. Evolution toward decreased human involvement
-

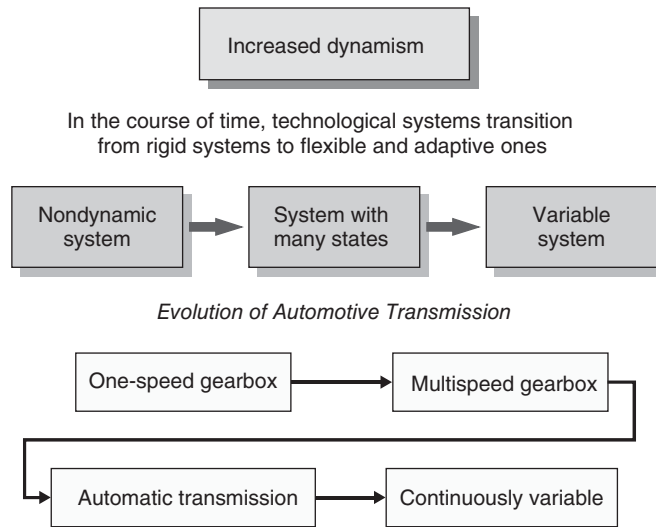
**Figure 11** Sublines of law 4, increased dynamism.

Figure 11 shows a few of the sublines of law 4, *increased dynamism*.

An analogy can be drawn between use of the laws of evolution and laws of motion. If the position of a moving object is known at a certain moment of time, any future position can be determined by solving equations containing velocity and direction. The laws of evolution serve as “equations” describing how the system will change as it travels through time. If the current position of the system is known, future derivatives can be “calculated” using the laws to indicate future positions. The implications to research-and-development initiatives, protection of intellectual assets, technology development strategy, patent strategies, and product development scenarios are profound.

5.6 Analytical Tools

In addition to the knowledge-based tools, Altshuller developed several analytical tools. The two most widely used are substance field modeling (Su-Field) and ARIZ.

5.7 Su-Field

The standard minimum system and its transformations (a generic formulation, according to the corollaries associated with Su-Field analysis) became the foundation of a set of standard solutions (76 standard solutions) that is effectively utilized for manipulation, with the intent of the model transformations analogically resulting in solutions to a specific problem. These solutions, or standard transformations, are grouped into five classes:

Class 1. Composition and decomposition of Su-Field models (SFM)

Group 1-1: Synthesis of a SFM

Group 1-2: Decomposition of SFMs

Class 2. Evolution of SFMs

Group 2-1: Transition to complex SFMs

Group 2-2: Evolution of SFM

Group 2-3: Evolution by coordinating rhythms

Group 2-4: Ferromagnetic SFMs (feSFMs)

Class 3. Transitions to supersystem and microlevel

Group 3-1: Transitions to bisystem and polysystem

Group 3-2: Transition to microlevel

Class 4. Measurement and detection standards

Group 4-1: Instead of measurement and detection—system change

Group 4-2: Synthesis of a measurement system

Group 4-3: Enhancement of measurement systems

Group 4-4: Transition to ferromagnetic measurement systems

Group 4-5: Evolution of measurement systems

Class 5. Special rules of application

Group 5-1: Substance introduction

Group 5-2: Introduction of fields

Group 5-3: Use of phase transitions

Group 5-4: Physical effects use

Group 5-5: Substance particles obtaining

Su-Field analysis

6 PROBLEMS WITHOUT CONTRADICTIONS

Overcoming contradictions solves both simple and complex problems. Why do contradictions occur? Because, striving to improve the world around us, the inventor demands a lot from technical objects. This is logical, for in order to meet the increasing demand, technical systems (TSs) should constantly increase in efficiency (or decrease in harmful, or redundant, properties). This means that one group of inventive problems focuses on improving the existing technical systems. Once involved in the technological evolution process, they start facing contradictions. The increasing demand cannot always be met by improving the existing TS. This gives rise to a question: Are there problems where no contradiction can be defined?

Example. In the course of reconstruction, a match factory was equipped with high-performance machines that doubled the factory's production rate. Yet, there was an operation that slowed down the whole process: packing the ready matches into boxes. The old machines could not cope with twice as much production. Lack of space made it impossible to install two packing lines. Finally, a decision was made to remove the out-of-date packing equipment. The old equipment had some deficiencies, too. It was "blind" and would often pack reject matches without heads or pack the wrong number of matches. Therefore, it became urgent to find an accurate method for packing millions of matches into boxes. There was a requirement for a system that would detect faulty matches.

There was no visible contradiction in this problem, but still there was the need to find a solution. The introduction of a small amount of ferromagnetic powder (application of a standard form Class 4, Group 4-4) to the ignition compound gave slight magnetic properties to each match. This was enough to orient the matches in a magnetic field and pack them faster and with higher accuracy (for a magnet of certain square surface attracts a fixed number of matches).

Let us analyze the problem and its solution in detail. First, as the conditions of the problem suggest, there was nothing to improve. The old TS was dismantled; therefore, a new system should be created. The matches were there, but what were we supposed to do with them? Should we count, orient, or package the matches? The problem was solved using the introduction of a ferromagnetic powder into the ignition compound of the match heads and using a magnetic field to create a system that could easily detect and control defect reduction in the packaged system.

In the beginning, there was one substance (the matches, S_1), and in the end there were two substances (the matches, S_1 , and the ferromagnetic powder, S_2) and one field (magnetic, F_M). The system is depicted in Fig. 12.

How does the system work? The magnetic field (F_M) acts on the ferromagnetic powder, S_2 , which in turn acts on the matches (S_1). Graphically the operation can be represented as depicted in Fig. 13.

In other words, one should work from a single element (S_1) toward a system of interacting elements (S_1 , S_2 , and F_M). A double arrow (to avoid confusion with arrows that indicate the interaction between elements) indicates this transition. The entire process of transition is displayed in Fig. 14.

All this resembles the symbolic representations of a chemical reaction. Two elements (e.g., oxygen and nitrogen) are heated (i.e., an external thermal field is introduced). As a result of interaction, they form a molecule of water; but, if a single atom is withdrawn from the molecule, the water will disappear. Can we treat the right-hand triangle of this technical reaction, in

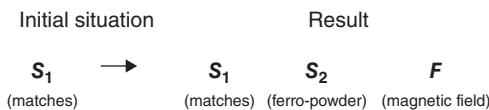


Figure 12 Pre-Su-Field analysis.

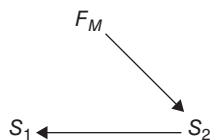


Figure 13 Su-Field model for example system.

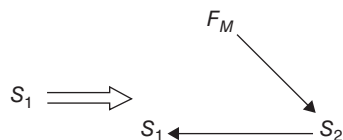


Figure 14 Incomplete system and transformation to solution model using Class 4, Group 4.4, from 76 standard solutions.

Fig. 14, as a “molecule” of a TS? Let us validate this idea: Will the system work if we withdraw any of the substances? No, the system will fall apart and cease to be a system. The same holds true for the situation in which the field is withdrawn. Does this mean that the system’s operation is secured by the presence of all three of the elements? Yes. This follows from the main principle of materialism: A substance can only be modified by material factors, i.e., by matter or energy (a field). With respect to a TS, this principle is as follows: A substance can only be modified as a result of a direct action performed by another substance (e.g., impact—mechanical field) or by a field action of another substance (e.g., magnetic) or by an external field. As a consequence, the minimal number of elements any TS consists of is three: two substances and a field—thus the concept of a minimal TS was named a Su-Field.

7 RULES FOR THE INVENTOR: SU-FIELD SYNTHESIS

Discarding redundancies, SFMs shed light on the essence of transformations (synthesis and evolution) of technical systems and allow the use of universal technical language to represent the process of solving any inventive problem. That is why analysis of Su-Field structures in those parts of technical systems where contradictions occur under transformation is called Su-Field analysis. Su-Field analysis presents a general formula that shows the direction of solving the problem. This direction depends heavily on the initial conditions of the problem. Consider the example problem: Any slightest alteration of conditions will profoundly change the process of solving the problem. For example, no materials may be introduced into the match head, no cooling medium can be poured into the hollow boom of the robot, etc. How can you decide which step to take?

The SFM is defined as follows: A Su-Field model is a representation of the minimal, functioning, and controllable technical system.

Quite often, conditions contain two substances and a field that have insufficient interaction and cannot be replaced with other substances or field. That is, the SFM is there (all three elements are present) and, at the same time, it is not there. It simply will not work. The same may happen after completing a SFM. That means that the SFM needs to be improved: The substances should become controllable, the field should have a desired effect, and the character of interaction of elements should proceed as required. There is a set of transformation rules for substances and fields in SFMs. The following is one such rule (see also Fig. 15):

Formation of complex Su-Field by introducing an easily controllable admixture possessing desirable properties into the substance. The admixture can be introduced into the substance (internal complex Su-Field) or, where internal introduction is inadmissible, placed outside the substance (external complex Su-Field).

- a. *Internal Complex Su-Field.* Wetting of fabric; foaming of varnish (problem 3); emergence of multicolored inserts impressed at certain distance to the cutting edge indicates the wear of the cutting tool (Soviet patent 905,417).
- b. *External Complex Su-Field.* Admixing ferromagnetic powder to cereal, production of hollow metal porous balls: Polystyrene balls are given a metal coat and

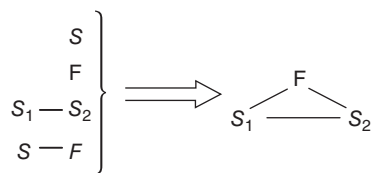


Figure 15 Transformational rules for SFM.

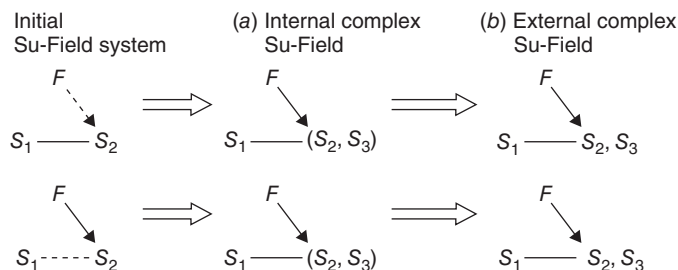


Figure 16 Complex SFM. Nonexistent interactions shown by dashed lines; parentheses indicate internal.

subsequently dissolved in organic solvent (U.S. patent 3,371,405). To avoid rumpling, the corrugations of the thin surface are filled with low-melting-point metal, which is withdrawn after treatment (Soviet patent 776,719) (see Fig. 16.)

8 CLASS 4: MEASUREMENT AND DETECTION STANDARDS

Group 4-1: Instead of Measurement and Detection—System Change

Standard 4-1-1. If we are given the problem of detection or measurement, it is proposed to change it such that there should be no need to perform detection or measurement at all.

EXAMPLE. To prevent a permanent electric motor from overheating, its temperature is measured by a temperature sensor. If the poles of the motor are made from an alloy with a Curie point equal to the critical value of the temperature, the motor will stop itself.

Standard 4-1-2. If we are given the problem of detection or measurement and it is impossible to change the problem to remove the need for detection or measurement, it is proposed to replace direct operations on the object with operations on its copy or picture.

EXAMPLE. It might be dangerous to measure the length of a snake. It is safe to measure its length on a photographic image of the snake and then recalculate the obtained result.

Standard 4-1-3. If we are given the problem of measurement and the problem cannot be changed to remove the need for measurement and it is impossible to use copies or pictures, it is proposed to transform this problem into a problem of successive detection of changes.

Note: Any measurement is carried out with a certain degree of accuracy. Therefore, even if the problem deals with continuous measurement, one can always single out a simple act

of measurement involving two successive detections. This makes the problem considerably simpler.

EXAMPLE. To measure a temperature, it is possible to use a material that changes its color depending on the current value of the temperature. Alternatively, several materials can be used to indicate different temperatures.

Group 4-2: Synthesis of Measurement Systems

Standard 4-2-1. If a non-SFM is not easy to detect or measure, the problem is solved by synthesizing a simple or dual SFM with a field at the output. Instead of direct measurement or detection of a parameter, another parameter identified with the field is measured or detected. Refer to Fig. 17.

If the conditions contain limitations on the introduction or attachment of substances, the problem has to be solved by synthesizing a Su-Field model using external environment as the substance: S_{se} is the substance from the surrounding environment. The left part of the formula coincides with that in the previous formulas.

EXAMPLE. To detect a moment when a liquid starts to boil, an electrical current is passed through the liquid. During boiling, air bubbles are formed; they dramatically reduce electrical resistance of the liquid.

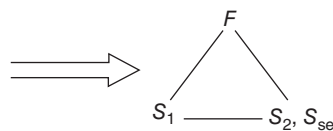
Standard 4-2-2. If a system (or its part) does not provide detection or measurement, the problem is solved by transition to an internal or external complex measuring SFM, introducing easily detectable additives.

EXAMPLE. To detect leakage in a refrigerator, a cooling agent is mixed with a luminophore powder.

Standard 4-2-3. If a system is difficult to detect or to measure at a given moment of time and it is impossible to introduce additives in the object, then additives that create an easily detectable and measured field should be introduced in the external environment and the changing state of the environment will provide an indication of the state of the object.

EXAMPLE. To detect the wear of a rotating metal disc in contact with another disk, it is proposed to introduce luminophore into the oil lubricant, which already exists in the system. Metal particles collecting in the oil will reduce luminosity of the oil.

If the conditions contain limitations on the introduction or attachment of substances, the problem has to be solved by synthesizing a Su-Field model using external environment as the substance:



S_{se} is the substance from the surrounding environment. The left part of the formula coincides with that in the previous formulas.

Figure 17 Synthesizing SFM using external environment as the substance.

Standard 4-2-4. If it is impossible to introduce easily detectable additives in the external environment, they can be obtained in the environment itself, e.g., by decomposing it or by changing the aggregate state of the environment.

Note: Specifically, gas or vapor bubbles produced by electrolysis, cavitation, or any other method are often used as additives obtained by decomposing the external environment.

EXAMPLE. The speed of water flow in a pipe might be measured by the amount of air bubbles resulting from cavitation.

Group 4-3: Enhancement of Measurement Systems

Standard 4-3-1. Efficiency of a measuring SFM is enhanced by the use of physical effects.

EXAMPLE. The temperature of liquid media can be determined by measuring the change in the coefficient of retraction, which depends on the value of the temperature.

Standard 4-3-2. If it is impossible to detect or measure directly the changes that take place and if no field can be passed through the system, the problem is to be solved by exciting resonance oscillations (of the whole system or of its part), whose frequency change is an indication of the changes that take place. Refer to Figs. 18 and 19.

EXAMPLE. To measure the mass of a substance in a container, the container is subjected to mechanically forced resonance oscillations. The frequency of the oscillations depends upon the mass of the system.

Standard 4-3-3. If no resonance oscillations can be excited in a system, its state can be determined by a change in the natural frequency of the object (external environment) connected with the system under control.

EXAMPLE. The mass of boiling liquid can be determined by measuring the natural frequency of gas resulting from evaporation.

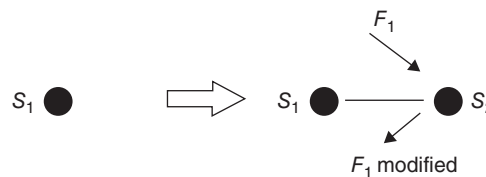


Figure 18

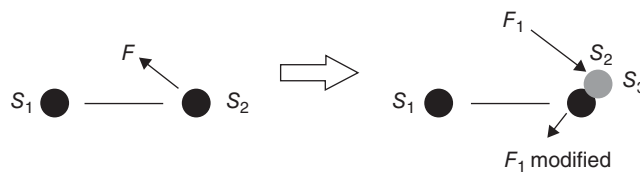


Figure 19

Group 4-4: Transition to Ferromagnetic Measurement Systems

Standard 4-4-1. The efficiency of a measuring SFM is enhanced by using a ferromagnetic substance and a magnetic field.

Note: The standard indicates the use of a ferromagnetic substance that is not crushed.

EXAMPLE. A group of students developed a method of measuring speed, direction, time, and operating status of an operating system designed to unwind some type of material from one spool to another spool. To take mechanical rotations and put them in the form of analog pulses that could be analyzed by either a microprocessor or electronic component through a pulsed tachometer, the following detection method was developed. A pulsed tachometer can detect rotations of a rotating shaft that contain a ferromagnetic rotor comprised of “iron brushes” perpendicular to the axis. The magnet in the pickup sensor creates a magnetic field around the sensor. When the iron brushes on the rotor pass through the magnetic field, the flux change induces an electromotive force (EMF) in a coil sensor. These create analog pulses that can be used to determine operating speed, time, direction, and status.

Standard 4-4-2. Efficiency of detection or measurement is enhanced by transition to feSFM, replacing one of the substances with ferromagnetic particles (or adding ferromagnetic particles), and by detecting or measuring the magnetic field.

EXAMPLE. In an effort to orient or align numerous objects, ferromagnetic material can be added to the same portion of each object to be aligned. A magnet can then be used to attract the ferromagnetic portion of the object, thus orienting or aligning the objects.

Standard 4-4-3. If it is required to raise a system’s efficiency of detection or measurement by going to a feSFM, while replacement of the substance with ferromagnetic particles is not allowed, the transition to the feSFM is performed by building a complex feSFM, introducing (or attaching) ferromagnetic additives to the substance.

EXAMPLE. The addition of iron oxide (a ferromagnetic powder) is now included as a pigment in black ink to validate currency and other negotiable documents. This technology is in continual development as computers and high-quality color printers make counterfeiting an elementary process. The magnetic fields from these particles produce signatures that, when read by magnetic sensors, can also be used to determine denominations of currency by vending or change machines.

Standard 4-4-4. If it is required to enhance a system’s efficiency of detection or measurement by going over to a feSFM, while introduction of ferromagnetic particles is not allowed, ferromagnetic particles are to be introduced in the external environment.

EXAMPLE. The discovery of the electron resulted in extreme advances in the chemistry field. In 1927, Wolfgang Pauli developed a formal representation of the electron spin concept. Experimentation in 1967 produced data that indicated that electrons from ferromagnetic particles (Fe, Co, and Ni) were not spin polarized as had been previously theorized. To continue testing, an ultrahigh vacuum was constructed where photoemissions of electrons could be performed down to 4.2 K and in magnetic fields up to 50 kOe. This device obtained strikingly different results: The electrons photoemitted from various particles were highly spin polarized. Continued research allowed for the development of spin polarization spectroscopy, helping scientists to further understand magnetism. Recent testing utilizing thin ferromagnetic films indicates that the films may be useful in acting as a spin filter similar to plastic foils used with polarized light.

Standard 4-4-5. Efficiency of a feSFM measuring system is enhanced by the use of physical effects, such as going through the Curie point, Hopkins and Barkhausen effects, magnetoelastic effect, etc.

EXAMPLE. Diagnosing and forecasting residual life of steel structures are important in determining the safety of large structures. Material magnetic memory (MMM) is effective in the assessment of the stressed–strained state of structures. This method envelops the theory that in zones of stress and strain concentration there are irreversible changes of the magnetic state of ferromagnetic items. Change of residual magnetization in tension, compression, torsion, and cyclic loading of ferromagnetic items is directly related to the maximal acting stress. The operator moves a sensor measuring the residual magnetic field intensity (H_p , A/m) along the weld over the entire perimeter and then transversely to the weld with the amplitude of deviation from the weld edge for 30–50 mm toward the base metal of the pipe element. The second operator records in the log book the data on residual magnetization of the metal, namely magnetic field intensity with the plus or minus sign. An abrupt change of the sign and value of H_p points to a concentration of residual stresses along the $H_p = 0$ line for a specific section of the welded joint. The main purpose of MMM is detection of the most critical sections and components in the controlled plant, which are characterized by strain concentration zones. After MMM, the traditional methods of nondestructive testing (ultrasonic test, X-ray, and eddy current inspection, etc.) are used to determine the presence of a particular defect.

Group 4-5: Evolution of Measurement Systems

Standard 4-5-1. Efficiency of a measuring system at any stage of its development is enhanced by transitioning to a measuring bi- or polysystem.

Note: For a simple formation of bi- and polysystems two or more elements are to be combined. The elements to be combined may be substances, fields, Su-Field pairs, and whole SFMs.

EXAMPLE. It is difficult to accurately measure the temperature of a small beetle. However, if there are many beetles put together, the temperature can be measured easily.

Standard 4-5-2. Measuring systems are developed toward a transition to measuring the derivatives of the function under control. The transition is performed along the following line:

Measurement of a function

→ measurement of first derivative of the function

→ measurement of second derivative of the function

EXAMPLE. Changes of stress in the rock are defined by the speed of changing the electrical resistance of the rock.

9 ALGORITHM FOR INVENTIVE PROBLEM SOLVING

ARIZ is the primary problem-solving tool in TRIZ. ARIZ was published in 1959 and revised many times: ARIZ-61, ARIZ-64, ARIZ-65, ARIZ-71, and ARIZ-85. Each revision improved the structure, language, and length of the algorithm. In its current state, we have a carefully crafted set of logical statements that transform a vaguely defined problem into an articulation of one with a clearly defined number of contradictions.

The assumptions designed into ARIZ are that the true nature of the problem is unknown and the process of finding a solution will follow the problem solver's vector of psychological inertia. It is why many of the steps in ARIZ are reformulations of the problem. With each reformulation, the problem is viewed from a different vantage point, yielding the possibility of new and novel ideas.

In mathematics, an algorithm is a precise set of steps designed to arrive at a single outcome. No consideration is given to the personality of the problem solver or to any changeable external conditions. The process is rote. In a broader context, an algorithm is a process following a set of sequential steps. ARIZ falls within that broader definition. ARIZ is a structured set of logic statements that guide the process of invention through a series of formulations and reformulations of the problem. If a chronic technological problem persists even after many attempts to solve it, the reason is often because the wrong problem is being solved. The selection of which problem to solve in an inventive situation is the starting point. It is critical that this selection is correct if there is any hope of arriving at an inventive solution in a timely manner.

As with any systematized process, ARIZ is dependent on the innate intelligence and knowledge of the subject matter expert and the skill with which he or she utilizes the tool. The strength of ARIZ, however, is that the process of thinking inventively is stripped of psychological inertia and regulated in a stepwise fashion toward the ideal solution, or in TRIZ terms, the IFR. The result is that the innate knowledge of the inventor is leveraged so that he or she is forced into thinking "inventively," e.g., into the solution space containing the most inventive ideas. Once the person is in the solution space, there are a number of inventive principles, analogs or Su-Field models that promote "thinking outside the box." See Fig. 20.

9.1 Steps in ARIZ

The architecture of ARIZ is composed of three major processes that are subdivided into nine high-level steps, each with their own substeps. ARIZ is designed to utilize all of the tools in TRIZ, including:

- Ideality
- The ideal final result
- Elimination of physical and technical contradictions
- Maximal utilization of the resources of the system
- SFMs and standard solutions
- The 40 inventive principles

ARIZ is designed to manage the inventive process on two types of problems: micro- and macro-problems. A microproblem is focused on solving a contradiction contained within the system while a macroproblem is a redesign of the entire system. ARIZ is iterative in that the inventor is provided several alternative paths to solving a problem. If all the solutions generated at the microlevel are unsatisfactory, the problem must be solved at the macrolevel.

A portion of the algorithm (Stage 1—Formulation of the problem) is detailed below.

1. Problem analysis

1.1. Microproblem. Write down the conditions of the microproblem (*do not use technology-specific jargon*):

- A technological system for (*specify the purpose of the system*) that includes (*a list of the main elements of the system*). Technical contradiction 1: (*formulate*).
- Technical contradiction 2: (*formulate*).

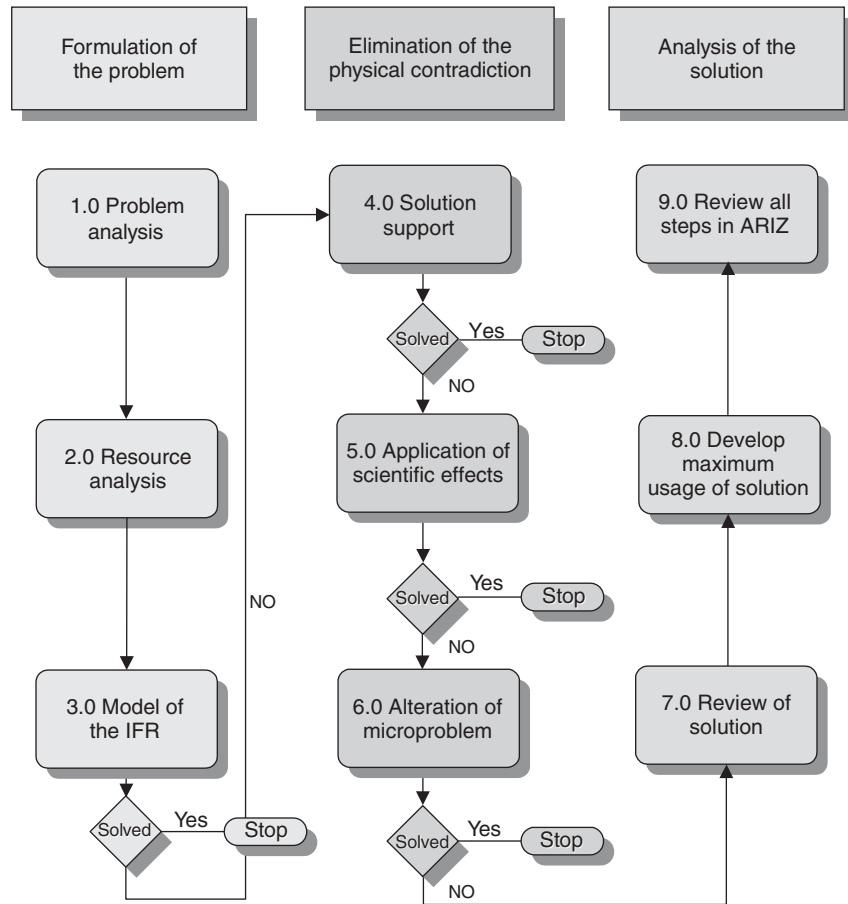


Figure 20 ARIZ flowchart.

- It is required to achieve (*specify desirable result*) without incurring (*specify the undesirable result*) with minimal changes or complications introduced into the system.

Note: Technical contradictions are defined using nouns for the elements in the system and actionable verbs describing the interaction between them.

1.2. Conflicting elements. Identify the conflicting elements: An article and a tool
Rules:

1. If an element can be in two states, point out both of them.
2. An article is an element that is to be processed or improved. A tool is an element that has an immediate interaction with the article.
3. If there is more than one pair of the identical conflicting elements, it is sufficient to analyze just one pair.

1.3. Conflict intensification. Formulate the intensified technical contradiction (ITC) by showing an extreme state of the elements.

1.4. Conflict diagrams. Compile diagrams of the intensified technical contradictions.

1.5. Selection of the conflict. Select from two conflict diagrams, one for further analysis.

Rules:

1. Select a diagram which better emphasizes the main (primary) function.
2. If intensification of the conflicts resulted in impossibility of performing the main function, select a diagram that is associated with an absent tool.
3. If intensification of the conflicts resulted in elimination of the article, use a "95% principle."
4. Select a diagram which better emphasizes the main function but reformulate an associated technical contradiction by showing, not extreme, but very close to extreme, states of the elements.

1.6. Model of solution. Develop a model of solution by specifying actions of an X-resource capable of resolving the selected ITC:

- It is required to find such an X-resource that would preserve (*specify the useful action*) while eliminating (*specify harmful action*) with minimal changes or complications introduced into the system.

1.7. Model of solution diagram. Construct a diagram of the model of the solution.

1.8. Substance field analysis. Compile a Su-Field diagram that models the solution: *Compile a SFM representing a selected ITC.*

- *Compile a desirable SFM illustrating resolution of the conflict.*
- *Select the appropriate standard solution and compile the complete Su-Field transformation.*

2. Resource analysis

2.1. Conflict domain. Define the space domain within which the conflict develops.

2.2. Operation time. Define the period of time within which the conflict should be overcome:

- Operation time is associated with the time resources available:
 1. Preconflict time T_1
 2. Conflict time T_2
 3. Postconflict time T_3
- It is always preferable to overcome a conflict during T_1 and/or T_2

2.3. Substance and energy resources. List the substance and energy resources of the system and its environment:

- The substance and energy resources are physical substances and fields that can be obtained or produced easily within the system or its environment. These resources can be of three types:
 1. In-system resources:
 - a. Resources of the tool
 - b. Resources of the article
 2. Environmental resources:
 - a. Resources of the environment that is specific to the system
 - b. General resources that are natural to any environment such as magnetic or gravitation fields of the earth
 3. Overall system resources
 - a. Side products: waste products of any system or any inexpensive or free foreign objects

3. Model of ideal solution

3.1. Selection of the X-resource. Select one of the resources from 2.3 for further modification. *Rules:*

1. Select in-system resources in the conflict domain first.
2. Modification of the tool is more preferable than modification of the article.

3.2. First ideal final result. The first IFR can be formulated as follows: The X-resource, without any complications or any harm to the system, terminates (*specify the undesirable action*) during the operation time within the conflict domain while providing the (*specify the useful action*).

3.3. Physical contradiction. Formulate a physical contradiction:

- To terminate (*specify the undesirable action*), the X-resource within the conflict domain and during the operation time must be (*specify the physical state P*).
- To provide (*specify the desirable action*), the X-resource within the conflict domain and during the operation time must be (*specify the opposite physical state $-P$*).

3.4. Elimination of physical contradiction macro. Use methods for elimination of physical contradictions:

- Separation of opposite physical properties in time
- Separation of opposite physical properties in space
- Separation of opposite physical properties between system and its components
- Separation of opposite properties upon conditions
- Combination of the above methods

Note: When applying the separation principles, use one or a combination of the following techniques:

- Separation in time:
 1. Think of ways to make the X-resource have property P before or after the conflict and property $-P$ during the conflict.
 2. Use “high-speed” processes.
 3. Explore various phenomena possible for the X-resource developed during phase transitions.
 4. Change the parameters or characteristics of the X-resource using a field.
 5. Explore using phenomena associated with decomposition of the X-resource into its basic elementary structure and then its recovery, e.g., ionization, recombination, dissociation, association, etc.
- Separation in space:
 1. Divide the X-resource into two parts having properties P and $-P$ with one part in the conflict domain and the other outside the conflict domain.
 2. Combine the X-resource with a void, porosity, foam, bubbles, etc.
 3. Combine X-resource with other resources.
 4. Combine X-resource with a derivative of another resource (e.g., hydrogen and oxygen is a derivative of water).
- Separation between the system and its components. Divide the X-resource into several components in a way that one component has property P while the other has property $-P$.
 1. Decompose the X-resource into elementary particles, granules, flexible rods, shells, etc.
 2. Explore using the phenomena associated with the decomposition of the X-resource into its base elements.

10 CAVEAT

ARIZ is a highly developed complex tool and should not be used on typical straightforward engineering problems. Also, becoming proficient with ARIZ takes time and practice. As a general rule of thumb, it is recommended that an individual solve 10 problems with ARIZ before they claim a layman's level of competency with the tool.

11 CONCLUSION

TRIZ is a powerful comprehensive problem-solving tool. It is the product of a massive analytical study of the output of the world's best inventors and most creative inventions. The fundamental underlying principle of TRIZ is ideality. The ideality principle holds that over time systems evolve to higher levels of functionality through the elimination of internal contradictions and the efficient utilization of available resources.

In time, the study of inventions by Altshuller and others yielded a number of knowledge-based and analytical tools. Knowledge-based tools include the contradiction matrix, the 40 inventive principles, and the laws of systems evolution. Analytical tools include Su-Field analysis and ARIZ.

Contradiction as a goal is a tough sell to American engineers. We rely on "trades." For the TRIZ practitioner finding a contradiction is the answer. If something has to be on/off, hot/cold, liquid/solid, magnetic/nonmagnetic, or any other dichotomy, that contradiction *is* the answer. We are lucky to have the founding efforts of Jim Kowalick, the sustained efforts of Ellen Domb, and the addition of M. Michael Slocum at the *TRIZ Journal* (www.trizjournal.com). The online journal has archives also online which are readily available. Google TRIZ and you will find out a great deal of information. Some of the best case studies are locked in the vaults. The 300+ cases at Ontro by Michael Slocum are delineated in the *TRIZ journal*.

A number of these engineering tools and initiatives work together. The reader may note several *connects* in various chapters of this document. Another recent tool is Design for Six Sigma (see Chapter 17). Where creativity or inductive reasoning is used, TRIZ may perform a positive service, especially those involving teams. A short list would include concurrent engineering, value engineering, ergonomic factors in design, processes, patents, total quality management, knowledge management, dimensional management, Six Sigma, and technical areas where teams are stalling or where a team needs to know if it is on the correct technical path or change is happening at the appropriate pace. Refer to the ABET certified course on systems engineering lecture notes of J. McMunigal.

Acknowledgments

A special "thank you" to John Opfell for collaboration utilizing engineering tools in the 1980s, including TRIZ starting in 1992 and for insights gleaned from direct Russian translation; Sam Brooks for long Sunday "chalk talks" on engineering tools and projects within Boeing featuring TRIZ and other engineering tools; Jeffrey A. Wolfe, Six Sigma black belt (BB); and Kelly R. McMunigal, Six Sigma yellow belt for technical support.

BIBLIOGRAPHY

- G. Altshuller, *Creativity as an Exact Science*, Gordon & Breach, New York, 1984.
- G. Altshuller, *The Innovation Algorithm*, Technical Innovation Center, Worcester, MA, 1999.
- S. Batchelor, "Solving the Problems of Particle Filled Fibers Using the TRIZ Methodology," *TRIZ J.*, October 1999, available: www.triz-journal.com.
- V. Bosse and J. E. McMunigal, book review of *Solving Problems with TRIZ (An Exercise Book)*, *TRIZ J.*, May 2004, available: www.triz-journal.com.

- J. Carr, "Analysis of a Problem: Clogging of a Multi-Drum Filter Used in a Textile Application," *TRIZ J.*, August 1999, available: www.triz-journal.com.
- R. Champa and R. Handley, *Brainware for the Strategist*, Strategy Partners International, Mission Viejo, CA, 2003.
- T. G. Clapp and B. A. Dickinson, "Design and Analysis of a Method for Monitoring Felled Seat Seam Characteristics Utilizing TRIZ Methods," *TRIZ J.*, December 1999, available: www.triz-journal.com.
- T. G. Clapp and M. S. Slocum, "Theory of Inventive Problem Solving Pedagogy in Engineering Education, Part I," *TRIZ J.*, November 1998, available: www.triz-journal.com.
- S. Gahide, "Smart Garment for Firefighters," *TRIZ J.*, June 1999, available: www.triz-journal.com.
- A. M. Gasanov, B. M. Gochman, A. P. Yefimochkin, S. M. Kokin, and A. G. Sopelnyak, *Birth of an Invention*, Interpraks, Moscow, 1995.
- N. Gibson, "The Determination of the Technological Maturity of Ultrasonic Welding," *TRIZ J.*, July 1999, available: www.triz-journal.com.
- D. Heath, "Addressing Salt Issues in Textile Dyeing Using an ISQ and ARIZ," *TRIZ J.*, January 2000, available: www.triz-journal.com.
- V. J. Khona, "Increasing Speed of Yarn Spinning," *TRIZ J.*, August 1999, available: www.triz-journal.com.
- B. Kunst and T. Class, "Automatic Boarding Machine Design Employing Quality Function Deployment, Theory of Inventive Problem Solving, and Solid Modeling," *TRIZ J.*, January 2000, available: www.trizjournal.com.
- J. E. McMunigal, "In Memory of Genrich Altshuller," *triz-viet nam*, January 1999, available: www.trizvietnam.com.
- J. E. McMunigal, notes from systems engineering course, California State University—Long Beach, Spring 2000.
- K. Rantanen and E. Domb, *Simplified TRIZ*, CRC Press, Boca Raton, FL, 2002.
- D. Raviv, "Introduction to Inventive Problem Solving in Engineering," *TRIZ J.*, March 1997, available: www.trizjournal.com.
- E. I. Rivin, "Use of the Theory of Inventive Problem Solving (TRIZ) In Design Curriculum," *Innovations in Engineering Education, 1996 ABET Annual Meeting Proceedings*, pp. 161–164; *TRIZ J.*, March 1997, available: www.triz-journal.com.
- M. Roberts, "B-cyclodextrin Molecules and Their Use in Breathable Barriers," *TRIZ J.*, November 1999, available: www.triz-journal.com.
- Y. Salamatov, *TRIZ: The Right Solution at the Right Time*, Insytec B.V., Hattem, Netherlands, 1999.
- S. D., Savransky, *Engineering of Creativity*, CRC Press, Boca Raton, FL, 2000.
- M. Slocum and J. E. McMunigal, "TRIZ and the Deconstruction of the Major World Philosophies," No. 17, Altshuller Institute for TRIZ Studies, 2003, available: www.aitriz.org/2003/ABSTRACTS.htm.
- J. Terninko, A. Zusman, and B. Zlotin, *Systematic Innovation*, CRC Press, Boca Raton, FL, 1998.
- S. Ungvari, *TRIZ Two Day Workshop Manual*, Strategic Product Innovations, Columbus, OH, 1998.
- S. Ungvari, *TRIZ Refresher Course*, Strategic Product Innovations, Columbus, OH, 1999.
- S. Ungvari, *TRIZ Problem Solving Guidebook*, Strategic Product Innovations, Columbus, OH, 1999.
- S. Vijayakumar, "Maturity Mapping of DVD Technology," *TRIZ J.*, September 1999, available: www.triz-journal.com.

CHAPTER 13

DATA EXCHANGE USING STEP

Martin Hardwick

Rensselaer Polytechnic Institute & STEP Tools, Inc.
Troy, New York

1	WHAT IS STEP?	391	6	STEP FOR BUILDING INFORMATION MANAGEMENT	395
2	STEP APPLICATION PROTOCOLS	391	7	NETWORKING STEP	396
3	STEP FOR LIFE CYCLE	392		REFERENCES	396
4	STEP FOR MODEL-BASED DEFINITION	392			
5	STEP FOR MODEL-BASED MANUFACTURING	394			

1 WHAT IS STEP?

In design and manufacturing, many systems are used to manage technical product data. Each system has its own data formats so when multiple systems are being used the same data has to be entered multiple times leading to redundancy and errors. Although repeated data entry is not unique to manufacturing, it is more significant because product data are complex and three dimensional (3D). The National Institute of Standards and Technology has estimated its costs the united States \$90 billion annually.¹

Over the years many solutions have been proposed. The most successful have been data exchange standards. The first ones were national and focused on geometric data exchange. They include the Standard d'Exchange et de Transfer (SET) in France, the Verband des Automobilindustrie FlächenSchnittstelle (VDAFS) in Germany, and the Initial Graphics Exchange Specification (IGES) in the United States. Later a grand unifying effort was started under the International Organization for Standardization (ISO) to produce one international standard for all aspects of technical product data and named STEP for the Standard for Product Model Data.² Today nearly every major CAD/CAM system has a STEP interface for reading and writing product data.

2 STEP APPLICATION PROTOCOLS

The ultimate goal of STEP is to define models for the entire product life cycle for all kinds of products. The initial goal was to enable the exchange of 3D models of parts and assemblies. The types of systems that can use such an exchange are shown at the top of Fig. 1. STEP was the first neutral data standard to enable solid model data exchange. After its lead other standards, both formal ones such as IGES and de facto ones such as 3DXML also began to support the exchange of solid models, but usually in a more limited infrastructure.

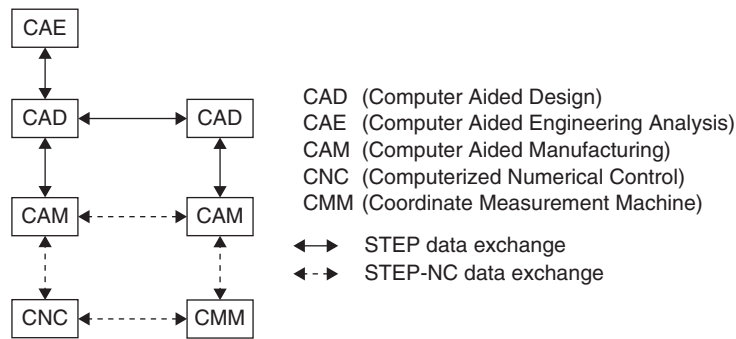


Figure 1 STEP and STEP-NC applications.

In 1996 there was a significant body of opinion that solid models could not be exchanged using a neutral data standard. In 1997 Ford, Allied Signal, and STEP Tools, Inc. demonstrated the technical feasibility in a prototype between two CAD systems. A pilot project, called AeroSTEP, was then organized by Boeing and its engine vendors to develop the first commercial translators. The goal was to enable the exchange of data about the interconnections between an aircraft engine and an airframe. The project started by exchanging simple faceted models and finished by exchanging complex assembly models.

As a result, vendor neutral implementation forums were formed in Europe, the Far East, and the United States and the quality of the translators was raised to a level that allowed anyone, including small organizations, to use STEP after about 2001.

3 STEP FOR LIFE CYCLE

In order to meet its goal of defining a model for the entire product life cycle STEP was divided into application protocols (APs) each defining a data exchange standard for a different type of product or a different stage of the cycle. Figure 2 is a list of the STEP APs as of October 2012.

The ability to support many protocols within one framework was a key strength of STEP. In the early years it allowed many groups to define protocols. All the protocols were built on the same set of integrated resources (IRs) so they all used common definitions for the same information. Each AP includes a scope, an activity diagram, and an application model describing the information requirements of those activities. These information requirements are then mapped into the common set of IRs and the result is a data exchange standard for the activities within the scope.

In the United States the most commonly implemented application protocol is AP-203. This protocol is used to exchange data describing designs represented as solid models and assemblies of solid models. In Europe a very similar protocol called AP-214 performs the same function. The common STEP architecture allowed the CAD, CAM, and CAE vendors to implement both with the same software library.

4 STEP FOR MODEL-BASED DEFINITION

Manufacturing needs many specifications to make a product. The most important is a description of the tolerances, but other kinds of specifications are also required, such as the material properties and the surface finishes. In STEP this information is known as product manufacturing information (PMI) and it drives the planning of the manufacturing processes and the selection of the manufacturing resources. In-house and out-house suppliers can then make the

Part 201	Explicit Drafting
Part 202	Associative Drafting
Part 203	Configuration Controlled Design
Part 204	Mechanical Design Using Boundary Representation
Part 205	Mechanical Design Using Surface Representation
Part 206	Mechanical Design Using Wireframe Representation
Part 207	Sheet Metal Dies and Blocks
Part 208	Life Cycle Product Change Process
Part 209	Design through Analysis of Composite and Metallic Structures
Part 210	Electronic Printed Circuit Assembly, Design and Manufacturing
Part 211	Electronics Test Diagnostics and Remanufacture
Part 212	Electrotechnical Plants
Part 213	Numerical Control Process Plans for Machined Parts
Part 214	Core Data for Automotive Mechanical Design Processes
Part 215	Ship Arrangement
Part 216	Ship Molded Forms
Part 217	Ship Piping
Part 218	Ship Structures
Part 219	Dimensional Inspection Process Planning for CMMs
Part 220	Printed Circuit Assembly Manufacturing Planning
Part 221	Functional Data and Schematic Representation for Process Plans
Part 222	Design Engineering to Manufacturing for Composite Structures
Part 223	Exchange of Design and Manufacturing DPD for Composites
Part 224	Mechanical Product Definition for Process Planning
Part 225	Structural Building Elements Using Explicit Shape Rep
Part 226	Shipbuilding Mechanical Systems
Part 227	Plant Spatial Configuration
Part 228	Building Services
Part 229	Design and Manufacturing Information for Forged Parts
Part 230	Building Structure Frame Steelwork
Part 231	Process Engineering Data
Part 232	Technical Data Packaging
Part 233	Systems Engineering Data Representation
Part 234	Ship Operational Logs, Records and Messages
Part 235	Materials Information for Products
Part 236	Furniture Product and Project
Part 237	Computational Fluid Dynamics
Part 238	Integrated CNC Machining
Part 239	Product Life Cycle Support
Part 240	Process Planning
Part 241	Building Life Cycle
Part 242	Managed Model Based Engineering

Figure 2 Step application protocols.

components and assemble the product. Each performs work for many customers so the PMI makes sure the deliverables have the right form, fit, and function for the final product.

During the course of the Industrial Revolution many types of tolerances were discovered and put onto drawings. Some of them may be less relevant in an age of automation, but for its first iteration STEP allows them all to be included as presentation tolerances or semantic tolerances or both. Presentation tolerances are the notes put onto drawings. They are difficult to process using intelligent software but they are presented for reasonably easy understanding by people. Semantic tolerances have a more precise mathematical definition. Usually they consist of a constraint applied to one or more surfaces with one or more plus/minus values referenced against datum planes. The two types are not mutually exclusive because the presentation tolerances can be used to present the semantic tolerances to the end user.

For model-based definition, STEP is developing a new AP as a unification of the best features of AP-203 and AP-214. The new protocol has the number AP-242 and the title Managed Model Based Engineering. Nearly all the CAD vendors have promised to upgrade their STEP translators to include AP-242 with presentation tolerances being implemented first. The U.S. Department of Defense is planning to require the delivery of these tolerances in future products as part of MIL Standard 31000. The semantic tolerances will take longer to implement, but they are a key attribute for the next stage of STEP deployment, which is predicted to be model-based manufacturing.

5 STEP FOR MODEL-BASED MANUFACTURING

STEP-NC AP-238 extends STEP to include manufacturing process and manufacturing resource information. CAM systems create these data and with STEP-NC they can send it to applications that intelligently control machining, material deposition, tape layup, and assembly operations.

A STEP-NC³ process is described as a series of operations that add or remove material. The volumes added or removed can be defined as features, generic shapes, or the implied result of tool movements. The operations are sequenced into working steps belonging to a work plan that has a geometric setup. Different types of work plans can be used to make the working steps conditional and concurrent. The assumed manufacturing machine and its associated tooling are also modeled as assemblies with kinematic movement and positional accuracy data. As a result a complete model of the as-designed process is sent to the shop floor, not just a set of codes that will work on one specific machine with one specific configuration (see the ISO 6983 “Gcode” standard).⁴

On the shop floor, systems can use AP-238 data to automate the adjustment of manufacturing parameters in response to changing manufacturing conditions: for example, to adapt to flexible fixtures, to enable the on-machine acceptance of parts, to make resource and performance optimizations, to enable last minute tooling changes, and to give better cost and schedule estimates to MES and ERP systems. Each application can be implemented as a custom CAM system or using other technology. The economic consequences include:

- If parts can be manufactured more easily and reliably, then there will be less need to keep physical copies in warehouses.
- If parts can be made independently of machine axis codes and other machine-dependent trivia, then the same program can be run on many different machines at many different sites.
- If parts include models of their in-process geometry and requirements, then manufacturing systems can be more concurrent and sophisticated with on-machine applications detecting and preventing errors.

Table 1 STEP-NC Testing

Phase	Demonstration Dates	Capabilities Shown	Purpose
1	November 2000 February 2002 January 2003 June 2003	Tool path generation from manufacturing features	Faster art-to-part
2	February 2005	CAM to CNC data exchange without postprocessors	CNC interoperability
3	May 2005 June 2006 July 2007	Integration of CAD GD&T data (as defined in AP-203 e2) with CAM process data (as defined in ISO 14649)	Integrated machining and measurement
4	December 2007 March 2008 October 2008	Cutting tool modeling (as defined in ISO 13399). Cutting cross-sectional modeling	Feed speed optimization
5	May 2009 September 2009 June 2010	Tool wear modeling Machine tool modeling	Resource management
6	October 2011 June 2012	Automated error compensation Accuracy prediction	Intelligent manufacturing

In the STEP framework, STEP-NC has the application protocol number AP-238. Table 1 shows the testing that has been performed to make sure the standard can support the new shop floor applications. The next stage of the testing is to demonstrate CAM to CAM data exchange to show how the new applications can be implemented.

6 STEP FOR BUILDING INFORMATION MANAGEMENT

In the last 10 years the program of STEP development has been reduced from the many APs shown in Fig. 2 to the two APs shown in Fig. 1. At the same time a new industry framework, based on the STEP technologies but with a different infrastructure, has been developed for the building construction industry known as industry foundation classes (IFCs).

IFCs were started by Autodesk as a set of class definitions in the C++ programming language. Like several similar programming language library specifications it was found to be too rigid and inflexible when extensions became necessary. Therefore, they looked at the STEP infrastructure and the way that it is designed to allow for many upward compatible extensions.

At the time, the building industries use of CAD and CAM was more limited than the aerospace and automotive industries. There are more engineers and more computer workstations in the building industry, but the mathematical modeling required is usually less sophisticated so it can be supported on lower powered systems and machines. This resulted in IFC making two changes to the STEP infrastructure. First they decided to use simplified geometries. This made the IFC models less geometrically accurate but more able to manage models of very large buildings.

Second they decided to amend the STEP integrated resources. They banned the use of multiple inheritance and replaced it with single inheritance. They also simplified the way that assemblies and properties are modeled while adding building specific requirements such as geospatial awareness. The model is now at its fourth iteration, and while it has more upward compatibility issues than STEP it has a large implementation base.

7 NETWORKING STEP

Despite the many successes of STEP there are still questions about the speed of its development and deployment.⁵ Many critics have pointed out that the Extensible Markup Language (XML) standards for e-commerce have been developed more quickly. Counter critics point out that STEP is used more frequently.

Fundamentally, product model data are different from other kinds of e-commerce data such as invoices, receipts, etc. The traditional method for communicating product model information is to make a drawing and the traditional method to communicate an invoice is to make a form. When you make a drawing or 3D model you need to define information with many subtle and complex relationships and this makes the STEP data exchange problem more difficult.

For 10 years the de facto path forward for STEP has been to develop e-commerce equivalents for its modeling language (EXPRESS) and its file format (Part 21). An initiative called PLCS for Product Life Cycle Systems took these developments to the logical conclusion and developed a series of XML schemas for different aspects of the product life cycle. To date, the response has been disappointing, perhaps because there are no revenues for a support industry, and perhaps because those developing data using XML are under more pressure to meet customer requirements than to conform to a standard.

STEP has important reasons to carry on using its Part 21 (ASCII) file format. Consequently, with the lack of momentum on switching, new interest is being placed on updating the existing technologies to support very large and distributed databases. A new edition of Part 21 is being prepared that will allow for interfile references using URL anchors and references. With this edition it will be possible for applications to concurrently work on different aspects of design, planning, and manufacturing.

REFERENCES

1. S. B. Brunnermeier and S. A. Martin, "Interoperability Cost Analysis of the U.S. Automotive Supply Chain," Research Triangle Institute, March 1999, available: <http://www.rti.org/publications/cer/7007-3-auto.pdf>.
2. "Industrial Automation Systems and Integration: Product Data Representation and Exchange—Overview and Fundamental Principles," ISO 10303-1:1994, International Organization for Standardization, Geneva, Switzerland, 1994.
3. "Industrial Automation Systems and Integration: Physical Device Control—Part 1: Overview and Fundamental Principles," Draft International Standard, ISO 14649-1:2001, International Organization for Standardization, Geneva, Switzerland, 2001.
4. S. H. Suh, J. H. Cho, and H. D. Hong, "On the Architecture of Intelligent STEP-Compliant CNC," *Int J. Computer Integrated Manufacturing*, **15**(2), 168–177, 2002.
5. M. Hardwick, "Third Generation STEP Systems That Aggregate Data for Machining and Other Applications," *Int J. Computer Integrated Manufacturing*, **23**(10), 893–904, 2010.

CHAPTER 14

ACHIEVING ENTERPRISE GOALS WITH NEW PROCESS TECHNOLOGY

Steve W. Tuszynski
Algoryx, Inc.
Los Angeles, California

1 INTRODUCTION	397	2.3 Tuszynski's Relational Algorithm	407
1.1 Historical Perspective on Technological Development	397	2.4 Material Selection	428
1.2 Traditional Approach	398	3 CONCLUSION	431
1.3 Problems with Traditional Approach	401	4 IMPLEMENTATION	431
1.4 Inefficiencies with Traditional Approach	406	APPENDIX A: TUSZYNSKI'S PROCESS LAW	433
1.5 Summary of Problems	406	APPENDIX B: DEFINITIONS	435
2 NEW TECHNOLOGY	406	APPENDIX C: NONTECHNICAL STATISTICAL GLOSSARY	436
2.1 History	406		
2.2 What New Technology Is Not	407		

1 INTRODUCTION

1.1 Historical Perspective on Technological Development

The main thrust of technological development has been to explore the relationships between causes and effects. We do this for many reasons, two of which are prime. The first reason is so that we can understand the natural processes that surround us and comprise our environment. The second reason is the basic premise that if we understand what the relationships are between causes and effects, we will be able to produce the effects we want by activating the causes that produce the results we want and minimizing the causes that detract from the results we want. This has been true from the earliest glimmerings of technology developed by mankind.

For example, what engineer has not learned Newton's law that $F = ma$? If we know the acceleration (a) that we want and the mass (m) of the item to be accelerated, then we can compute the force (F) required. If the mass of the item is fixed, then there will be only two variables and the acceleration will be proportional to the force applied. In this instance, if the force is doubled, the acceleration is doubled. If the force is quadrupled, so is the acceleration. What we learn from Newton is that acceleration is related to force. In this instance, the two variables are "co-related." We also say, with the same meaning, that they are "correlated."

When did we start thinking this way? Perhaps it was when our early ancestors first threw rocks at animals to defend themselves or get food or perhaps even earlier when early humans tried to figure out what kind of behavior it took to survive or reproduce. This approach is ingrained in the human mentality and has been the foundation on which we have built our

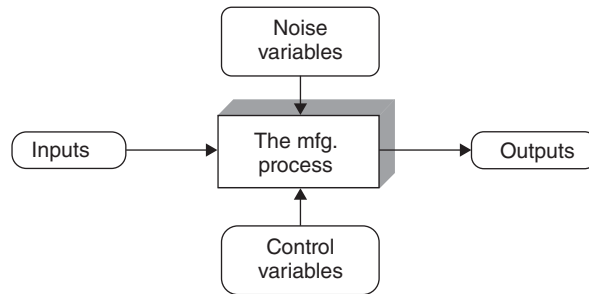


Figure 1 Traditional manufacturing flow diagram.

technology. This approach has been necessary, useful, and productive. It is hard to imagine our existence without our understanding the linkages between causes and effects.

Converting raw material to useful products in a manufacturing process involves converting inputs into outputs. More generally, the result of any natural process is change. The purpose of man-made processes is to produce change toward a desired goal or objective.

1.2 Traditional Approach

Manufacturing Process Flow Diagram

The balance of this chapter discusses process technology in the context of design, tooling, manufacturing, and quality engineering.* Figure 1 is a manufacturing process flow diagram. It shows inputs into and output from the manufacturing process. Control and noise variables influence the output for any given input.

Inputs

Inputs are items input into the process and can be physical or nonphysical. Inputs are usually physical items when the process produces physical output, but they can be nonphysical items where the process is an algorithmic or computational process. Nonphysical items can be of many types. In a manufacturing or simulation process, inputs are usually either variable or attribute data. Inputs can also be combinations of physical and nonphysical items.

The manufacturing process is a single step or a series of sequential steps that modifies the inputs to the process. Each stage in a manufacturing process will have an input and an output.† The input of any one stage will be the output of the preceding stage and the output of any one stage will be the input to the subsequent stage.

Control Variables

Control variables are the process parameters‡ controlled by an operator or process engineer. In essence, these are the knobs and dials, whether manually or automatically controlled, on the manufacturing equipment that are adjusted to produce conforming parts.§ Figure 2 shows

* Hopefully, those individuals schooled in sciences and technologies other than engineering and manufacturing will see applications from this chapter to their respective areas of expertise.

† For ease of reading, multiple inputs and / or outputs are referred to here as an input or an output.

‡ Pressures, temperatures, speeds, times, chemical concentrations, orientations, power settings, frequencies, intensities, agitation levels, etc., are typical process control settings.

§ Conforming parts meet engineering specification or drawing values. Conforming parts are defined as “good parts” and nonconforming parts are defined as “bad parts.”

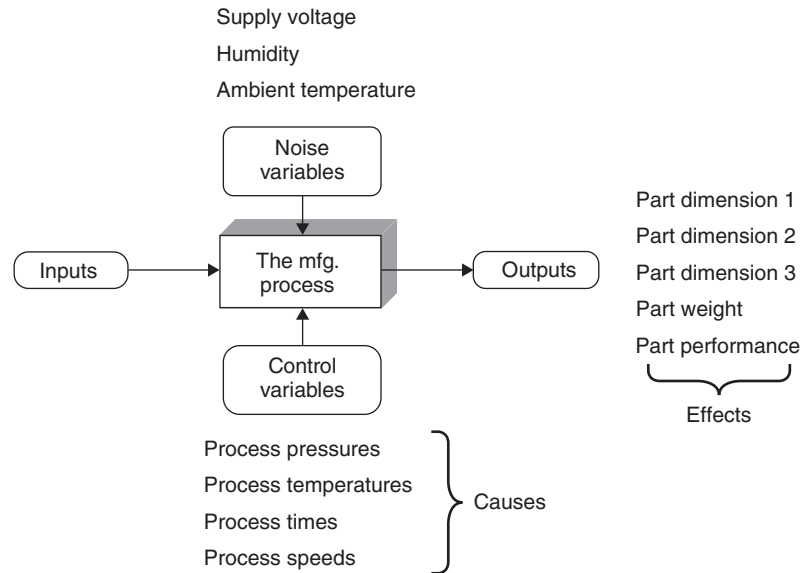


Figure 2 Process causes and effects.

typical process control variables such as pressures, temperatures, times, and speeds. These are the variables that are controlled by the process operator.

Noise Variables

Noise variables are those variables that influence the output of the process. Figure 2 shows various noise variables for a particular process. Noise variables are not controlled because either they cannot be controlled or we choose to not control them. Noise variables may be left uncontrolled for numerous reasons, including circumstances where:

- They are unknown.
- They are too expensive to control.
- They are too time-consuming to control.
- It is not possible to control them.
- Controlling them would not make an appreciable difference in the quality or producibility of the manufactured part.

Output Variables

The output of the process consists of the item to be manufactured or produced. The output of the process will have different characteristics. In the instance where parts are being manufactured, the characteristics will be referred to as part characteristics. Figure 2 shows various typical part characteristics such as dimensions, weight, or performance as outputs of the manufacturing process.

For purposes of this chapter, part characteristics are divided, arbitrarily and for convenience, into four categories:

1. Variable characteristics
2. Attribute characteristics

3. Material characteristics
4. Performance characteristics

Variable characteristics are most typically dimensions. Dimensions are usually subcategorized into critical and noncritical dimensions.*

Attribute characteristics are data describing a part characteristic not measurable on a number scale. Typical attribute characteristics are on or off, the presence or absence of undesirable characteristics, supplier A, B, or C, material type m, n, or o, etc. Some attributes, such as color, can be converted to variable data (for example, a combination of red, green, and blue) if it is worth the cost and effort. In the context of this chapter, visual attribute characteristics can be thought of as the presence or absence of some desirable or undesirable part characteristic as determined through visual inspection.

Material characteristics are the physical properties of the manufactured part. Typical material characteristics could be tensile strength, surface hardness, density, or reflectivity. Material characteristics are usually variable data. The categorization of part characteristics into these first three categories is not crucial but is a matter of convenience.

Performance characteristics refer to those characteristics that are measures of how well the part performs relative to its functional requirements. Performance characteristics are usually variable characteristics but can also be attribute characteristics.

Causes and Effects

Figure 2 identifies the control variables as causes and the output characteristics as effects. The foundation of the traditional approach is to relate causes and effects.

Traditional Approach to Cause and Effect

Figure 3 shows how the essence of the traditional approach is to determine the linkage between process causes and effects. This approach is premised on the logical belief that if we adequately understand the relationships between causes and the effects, then we should be able to set the control variables to values that produce the desired part characteristics. In essence, the traditional approach looks for the correlations that model the relationships between causes and effects.

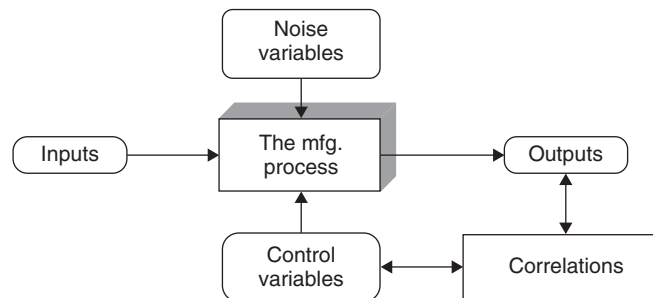


Figure 3 Traditional approach to cause and effect.

* Critical dimensions are those considered by the design engineer to be critical to form, fit, or function (performance).

Traditional Approach to Determining Correlations

The traditional approach to determining the correlations between causes and effects is to evaluate the part characteristics when parts are manufactured under different control settings (causal conditions). In some instances, the control settings will be deliberately changed to induce variation in the manufactured parts. In other instances, there may be enough natural variation in the control settings that, over time, enough data with enough variation will be generated so that the correlations can be determined.

Prior to the invention and application of efficient statistical methods, variation in part characteristics was usually induced by changing one control variable at a time and then determining how each part characteristic changed. Sequentially changing each control setting one at a time is very time and cost inefficient and can lead to erroneous conclusions.

Design of Experiments

Design of experiments (DOE) is a statistical methodology that has greatly improved the efficiency of determining the relationships between causes and effects. DOE is a large step forward in improving the time and cost efficiencies and in reducing erroneous conclusions. However, as discussed below, DOE does not universally explain all of the relationships between causes and effects.

1.3 Problems with Traditional Approach

As useful as the traditional approach is, there are many situations in which it is difficult, impossible, or uneconomical to determine the relationships between causes and effects. These situations can occur irrespective of whether DOE is used. This section discusses several situations that make determining the relationships between causes and effects impractical.

More Than a Few Control Variables—Many Relationships

Some processes have relatively few control variables, while others can have many. As shown in Fig. 4, plastic injection-molding processes, for example, can have over 20 control variables.

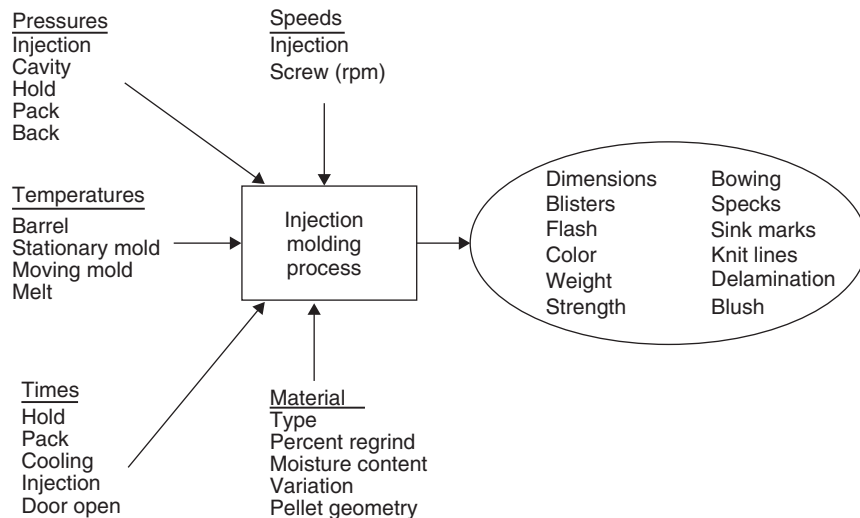


Figure 4 Typical plastic injection-molding variables.

When many control variables are involved, there are many cause-and-effect relationships to be evaluated and understood.

More Than a Few Control Variables—Many Control Setting Combinations

Further, as the number of control variables increases, the possible number of combinations of control variables increases geometrically. For most process variables, there are an infinite (analog) or large (digital) number of settings for each control variable. However, the situation can get complex even when there are only two or three levels chosen for each control variable. For example, if there are 20 control variables and each control variable is examined at only two settings—a high and a low value—then there are over one million possible combinations. If each of the 20 control variables is examined at three settings—a high, a nominal, and a low value—then there are over three billion possible combinations.

More Than a Few Part Characteristics

For products with multiple part characteristics, process engineers and operators have the difficult task of attempting to adjust process settings to produce parts with all dimensions simultaneously at target values. Some parts have a large number of critical characteristics. For example, the single plastic injection-molded part shown in Fig. 5 has 42 critical dimensions. In this instance, one must determine the relationship between each control variable and each of the 42 critical dimensions.

Different Responses to Control Variable Changes

Different part characteristics can have different responses to changes in control settings. Figure 6 shows how the length of a part increases when the process temperature setting is increased. However, the diameter of the part decreases as the process temperature setting is increased. One cannot increase both the length and diameter of the part by changing the process control setting.

Multiunit Processes

Some processes produce multiple parts for each process cycle. Injection-molded parts, for example, are frequently made with multicavity molds. Figure 7 illustrates this point for a part

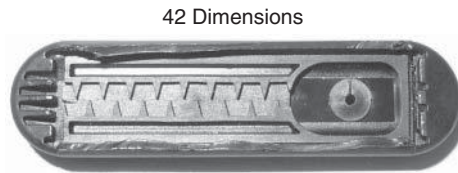


Figure 5 Typical plastic injection-molding variables: a single part with 42 critical dimensions.

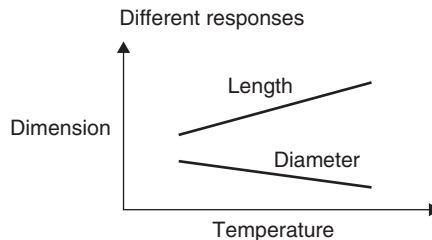


Figure 6 Responses can be different to changes in control settings.

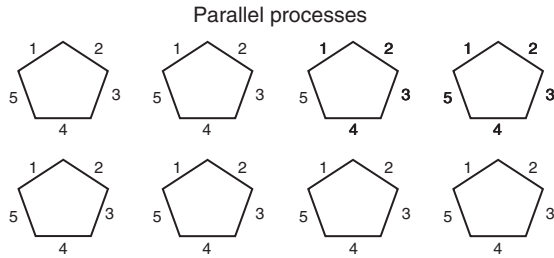


Figure 7 Multiunit processes greatly increase the number of part characteristics.

that has five critical dimensions and is manufactured in an 8-cavity mold. It is not unusual for molded plastic parts to be made with 8-, 16-, or 32-cavity molds. Molded rubber parts can be made with molds having hundreds of cavities. Semiconductor wafer fabrication processes can result in thousands of circuits on each wafer.

Multiunit processes can get quite complex when each part has many part characteristics. For example, when a part that has 42 critical dimensions is produced in a 32-cavity mold, each machine cycle produces 1344 separate critical dimensions.

Simple Interactions

Figure 8 illustrates a simple interaction between two control variables. In this example, if the pressure control variable is at setting level 1, the length of the part increases as the temperature control variable increases. However, if the pressure control variable is at setting level 2, the length of the part decreases as the temperature increases. Put more simply, the response of the length to changes in temperature depends on the value of the pressure. Simple interactions are common in many manufacturing processes.

Complex Interactions

Figure 9 illustrates a complex interaction between three control variables. In this instance, if the pressure control variable is at setting level 1, the length of the part:

- Decreases as the temperature increases when the speed is at level 1.
- Remains unchanged as the temperature increases when the speed is at level 2.
- Increases as the temperature increases when the speed is at level 3.

A different set of three response curves will also exist for the pressure control variable setting at level 2. Put more simply, the response of the length to changes in temperature depends not only on the value of the pressure but also on the value of the speed. Although complex

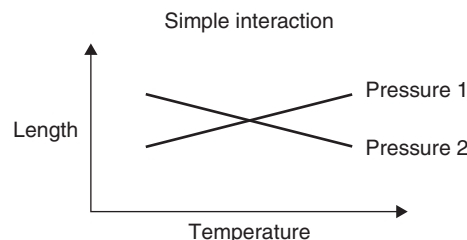


Figure 8 Simple interactions make it difficult to determine responses.

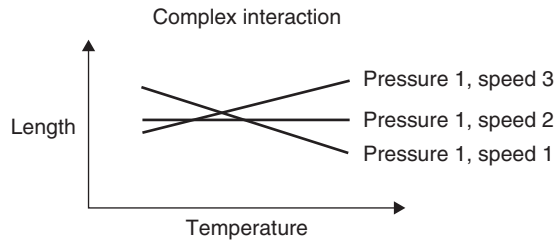


Figure 9 Complex interactions make it even more difficult to determine responses.

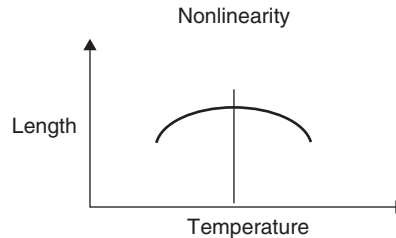


Figure 10 Nonlinear responses can lead to errors.

interactions are not as common as simple interactions, they do occur occasionally in manufacturing processes.

Nonlinear Responses

Figure 10 shows a nonlinear response between a part characteristic and a process control variable. The length increases, reaches a maximum, and then decreases as the temperature control variable increases. If one is sampling only two temperature levels to determine the response of the length to changes in temperature, one could conclude, depending on the two points chosen, that

- length increases as temperature increases,
- length is insensitive to changes in temperature, or
- length decreases as temperature increases.

Two types of errors can occur. The first is the linear approximation of a nonlinear response. The second is that a reversal, as noted immediately above, can occur, which would invalidate the conclusion on how much and in which direction to change the temperature.

DOE May Not Be Helpful

Design of experiments has proven to be a useful tool for circumstances where (1) one part or performance characteristic needs to be optimized and (2) there are few process complexities. When the first criterion is not met, DOE can and generally does give conflicting results. When the second criterion is not met, it is difficult to model the process and get useful results.

Small or Nonexistent Producibility Windows

The producibility window may be nonexistent, i.e., it can be impossible to produce good parts. Operators and process engineers can waste significant time learning this. Even if a set of

process settings can be found that produces good parts, the producibility window may be so small that even minor variations in process settings result in bad parts (ones that do not meet specifications).

Operator Technique

Small or nonexistent producibility windows frequently occur prior to achieving a good first article. This can make it difficult or impossible to produce good parts. The path of least resistance is for the operator to try to produce good parts by adjusting process settings.* Also, the techniques that worked for an operator in the past may not work for the current part.

Tolerance Relaxation Dependency on Operator Process Settings

When the producibility window is small or nonexistent, process engineers frequently ask the design engineer for tolerance relaxation on the problematic part characteristics (dimensions, for example). There is a potential problem with the method currently used to determine which design tolerances need to be relaxed and by how much. The problem occurs because the part characteristics clearly are dependent on the process settings selected by the operator. A subsequent change to the process settings can invalidate the need for the original tolerance relaxations and create the need for new tolerance relaxations on different part characteristics. Based on the parts produced by operator A, the process engineer could ask, for example, for an increase in the upper (+) tolerance on dimension X. Based on the parts produced by operator B (or operator A at a different point in time with different process settings), the process engineer could ask for a decrease in the lower (–) tolerance on dimension Y. No prior art technology has solved this problem or given the design engineer a ranking of the order in which to relax design tolerances or the size of the required tolerance relaxations independent of process settings.

Tooling Modification Dependency on Operator Process Settings

A similar but usually more grievous situation can occur with tooling, molds, and fixtures. A preproduction mold, for example, might go through five to eight different modifications before it is qualified to produce good parts. Tool and fixture modifications can be risky and time consuming. For all processes, dimensional results depend on the values selected for the process settings. This creates a problem for the tooling engineer when deciding how to modify the tooling or fixture. Use of different process settings can invalidate previously made tool, mold, or fixture changes. Tooling and fixture engineers refer to this as the “tyranny of the operator.” No prior art technology has solved this problem or, given the tooling engineer, in a single step and independent of process settings, the tooling modifications required to produce parts at design targets.

Trial and Error, Iteration, and Guesswork

The absence of a scientific method for eliminating the preceding complexities has resulted in trial-and-error, iteration, and guesswork attempts to produce good first articles and to produce the highest quality parts during production.

* Changing process settings is usually the first technique used to try to produce a good first article. If a good first article cannot be produced, the next step, usually after extensive “fiddling” by the process operator, is often to ask the design engineer for tolerance relaxation. If the design engineer does not relax the tolerances, the next option is usually to modify the tooling. Tooling modification is usually the last of these three options because of time, cost, and risk considerations. Finally, process variables can be monitored and controlled through various equipment.

Changing Process Technology

When tooling or fixtures cannot be adequately modified and/or tolerances cannot be adequately relaxed for a given material, and/or inadequate process control exists, then a more capable process must be used to produce the part. No prior art technology has enabled the design, tooling/fixture, process, and quality engineers to easily determine when this is the case.

1.4 Inefficiencies with Traditional Approach***Current Statistical Process Control Studies Are Inefficient***

During part development (first article, qualification, and certification), statistical process control (SPC) studies, when they are done, are usually done on all dimensions. During production, SPC studies, when they are done, are usually done on either all or a subset of dimensions. As will be shown later, this is inefficient and incurs unnecessary cost.

Current Process Capability Studies Are Inefficient

In a similar fashion, process capability studies are done on all dimensions. This is also inefficient and incurs unnecessary cost.

Shipping and Receiving Inspections Are Inefficient

In a similar fashion, shipping and receiving inspections are usually performed by sampling a subset of parts and measuring all critical dimensions. This incurs unnecessary costs for both the customer and supplier.

1.5 Summary of Problems

Historical gains in the quality of manufactured parts over the last two decades have been significantly eroded by increases in measurement and recording costs and by SPC and process capability (Cpk, Ppk) analysis costs. The efforts of engineers to modify preproduction tooling and fixtures to production tooling and fixtures are frustrated by operator changes to process settings. Process engineers have difficulty determining the values of process settings and design engineers find it difficult to design for producibility when there are multiple part characteristics. The use of standard design tolerances increases manufacturing costs. Determining tolerance relaxations is a trial-and-error process compounded by operator changes to process settings. Replacing obsolete materials can be problematic. The inability to produce parts at design target values decreases product performance.

Design, process, tooling, and quality parameters are interrelated because they co-jointly influence part characteristics and consequently how the part performs and decision making on whether or not the part is a conforming part. Prior art has not provided engineers and decision makers with an integrated system of technology that incorporates these interrelationships.

2 NEW TECHNOLOGY**2.1 History**

In 2000, a large, international original equipment manufacturer (OEM) was having difficulty producing good first articles for a new product line of injection-molded plastic parts. These problems stimulated the development of innovative technology that solved the problems. The new technology has been proven in several widely diverse industries* with numerous

* This new technology has been proven for plastic injection molding, sheet metal punching, sheet metal forming, CNC laser cutting, and semiconductor wafer fabrication. Case studies are underway or planned for plastic extrusion, rubber molding, hot and cold metal heading, plating, etching, and wire forming.

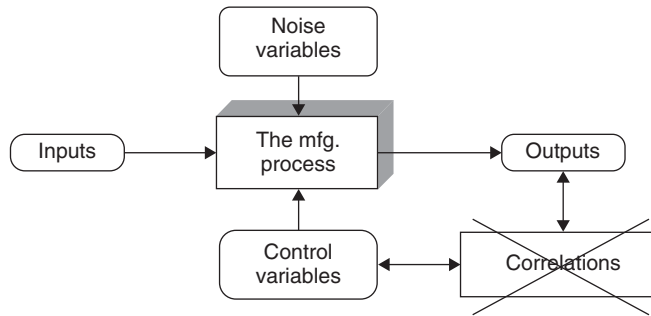


Figure 11 Traditional approach relates cause and effect.

case studies.[†] No changes are required to the manufacturing process. Huge reductions in measurement and analysis costs have been achieved, as well as increased quality, increased productivity, and reduced time to market. The bottom line has been significantly increased profits and return on investment (ROI). The new technology is called Tuszynski's relational algorithm (TRA).

2.2 What New Technology Is Not

It is sometimes easier to introduce new technology by stating it is not. Most importantly, as shown in Fig. 11, TRA does not attempt to determine the relationships (correlations) between causes (process settings) and effects (part characteristics), so TRA bypasses the process complexities and inefficiencies mentioned above.

The following are lists of what TRA is not and what it does not do. TRA is not any of the following types of computer programs:

- A SPC program
- A Cpk analysis program
- A DOE program
- A finite-element analysis (FEA) program
- A plastic flow simulation program

TRA is not used for:

- Designing tooling, molds, and fixtures
- Designing mold runner and gate systems

TRA does not use iterative procedures.

2.3 Tuszynski's Relational Algorithm

New Algorithms

Tuszynski's relational algorithm (TRA) is a pioneering system of interrelated algorithms that has led to breakthrough insights into manufacturing processes and to powerful new computational software (Fig. 12). This state-of-the-art technology reduces the cost of

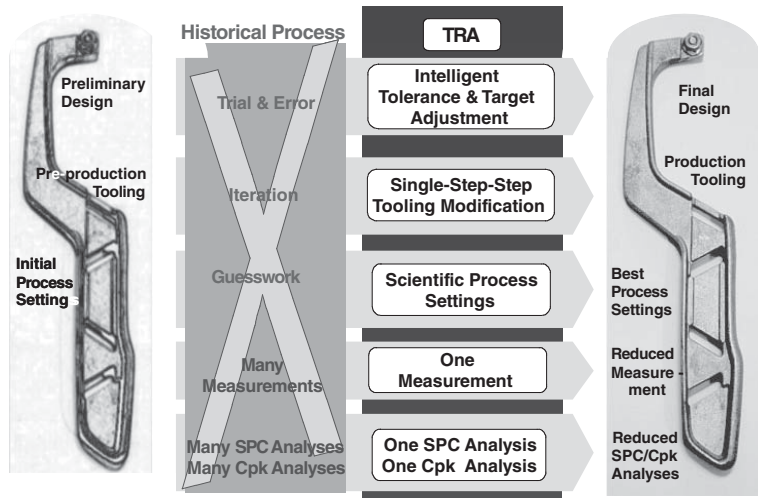
[†] The author would welcome feedback on applications of the algorithms presented in this chapter to new processes and industries.

Tuszynski's Relational Algorithm (TRA)

Use TRA™ to Achieve Lean and Six Sigma Goals

- Reduce Time-to-Market
- Improve Profits and ROI
- Maximize Quality
- Reduce Scrap, Rejects and Rework
- Eliminate Repeated Tooling and Mold Modifications
- Locate the Process Sweet Spot
- Simultaneously Optimize All Cpk's
- Optimize Design Targets for Producibility
- Optimize Design Tolerances for Producibility
- Simulate Tooling Changes without Time, Cost or Risk
- Improve CNC Programming
- Slash Inspection Costs
- Eliminate Destructive Inspection
- Slash SPC Analysis Costs
- Slash Cpk Analysis Costs
- Integrate Cross-Functional Decision-Making
- Improve Customer-Supplier Communication

Fastest Time to Market – Best Design, Best Tooling, Best Process, Least Analysis



ALGORIX, INC.™


 750 S. Bundy Drive, Ste. 304
 Los Angeles, CA 90049
 310-820-0987
www.algorix.com
steve@algorix.com

- Plastic Injection Molding
- Plastic Extrusion
- Rubber Molding

- Semiconductor Wafer Fab
- Thermoforming
- Cold and Hot Heading

- CNC Laser Cutting
- Sheet Metal Forming
- Sheet Metal Punching

© 2004-2005

Figure 12 Illustration of application of TRA across engineering functionalities.

enterprise quality management (EQM) operations and facilitates the implementation of lean strategies (LS) by achieving huge reductions in inspection, SPC, and process capability analysis (Cpk) costs, by providing optimized tooling, and by eliminating multiple redundancies in six sigma programs and quality operating systems (QOSs).

Foundation of New Algorithms

TRA is based on the fact that although the relationships between causes (process settings*) and effects (part characteristics†) may be difficult or impossible to determine, the relationships between effects for many processes are consistent and predictable irrespective of changes in the process settings. One of the part characteristics is selected as the predictor characteristic.‡ The predictor characteristic is the characteristic that is the statistically best predictor of all other part characteristics.

When Does TRA Work?

TRA works when there is correlation between part characteristics. This is generally true when the process adds or subtracts material or changes the shape or form of the material. If there is no correlation between part characteristics, TRA will not work.

Graphical Illustration of TRA

Figure 13 illustrates a condition where there are four interrelated dimensions on a part—*A*, *B*, *C*, and *P*—where *P* has been selected as the predictor dimension and *A*, *B*, and *C* are the predicted dimensions. The relationships between the dimensions are defined by regression lines fitted through data generated from manufactured parts.§

TRA Process Conclusions

Figure 13 is simple, yet sophisticated and complex at the same time. Figure 13 leads us to the following rather startling conclusion:

Even though the relationships between causes (process settings) and effects (part characteristics) may be difficult or impossible to determine, the relationships between effects are consistent and predictable irrespective of changes in the process settings!!!

This conclusion is startling from at least three perspectives. First, it does not matter what the complexities were or how many complexities were present when the data were generated. The relationships between part characteristics can be determined irrespective of the complexities. In essence, process complexities are eliminated. Second, the relationships between part characteristics are simple, understandable, and fixed.¶ The relationships can be easily visualized when they are presented graphically. Third, it does not matter which particular combination of process settings was used to determine any single point on any regression line.∥

* Pressures, temperatures, times, speeds, etc.

† Part characteristics include dimensions, weights, material characteristics, and performance characteristics.

‡ The predictor characteristic is referred to as the predictor dimension. Most applications of TRA to date have been with dimensional part characteristics.

§ The regression lines shown in Fig. 13 assume perfect correlation. The assumption of perfect correlation is used here to simplify the graphs and the discussion. It will be removed later in the chapter.

¶ The relationships are fixed as long as tooling is fixed and materials are unchanged. If the tooling dimensions change, the changes in relationships are predictable.

∥ For certain processes, it is possible to produce a part with a specified part characteristic (or with a specified set of part characteristics) by selecting different combinations of process settings.

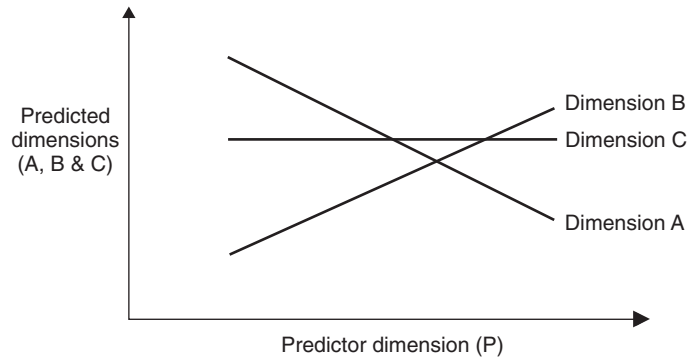


Figure 13 A part with four dimensions.

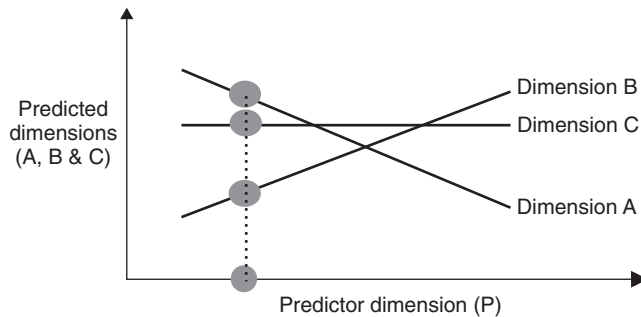


Figure 14 Regression lines constitute single-degree-of-freedom system.

The relationships between part characteristics are fixed in spite of the multiplicity of combinations of process settings that can result in a specific dimensional outcome.

The relationships shown in Fig. 13 represent the entire universe of possible part relationships!!!

All 3 billion (20 process settings at 3 levels) possible process setting combinations are encompassed by Fig. 13.

Single-Degree-of-Freedom System

The system of regression lines (and part characteristics) shown in Figs. 13 and 14 is also a single-degree-of-freedom system. Figure 14 shows that when the value of the predictor dimension (P) is known, the values for all predicted dimensions (A, B, and C) are also known.

When the value of the predictor characteristic is known, the value of all predicted characteristics is known.

In essence, the entire system of part dimensions is collapsed into a single predictor dimension. Instead of measuring all dimensions, only the predictor dimension needs to be measured. Instead of performing SPC analysis on all dimensions, SPC analysis needs to be performed only on the predictor dimension.* Instead of performing process capability (Cpk) analysis on

* Important restrictions may apply to multiunit processes. Refer to later discussions on in-control versus out-of-control conditions. Even with these restrictions, huge savings are still obtainable.

all dimensions, process capability analysis needs to be performed only on the predictor dimension. If destructive measurement is being done, there is potential for eliminating the destructive measurement by measuring only the predictor dimension. Any point on Figs. 13 and 14 can usually be obtained by alternative combinations of process settings, meaning that any given point in Figs. 13 and 14 is not restricted to one unique combination of process settings.

Operating Point Adjustments

The operating point is defined as a point on any of the regression lines. The operating point is adjusted by changing one or more process settings. This is analogous to sliding a bead along a wire.

Generating Input Data

Input data for TRA processing is obtained from:

1. Existing manufacturing data if there is adequate variation in the data
2. Existing DOE data
3. New data generated by inducing variation in the manufacturing process

In some cases, there may be sufficient existing manufacturing data to generate the correlation charts. In many companies, DOE studies exist on parts and manufacturing processes. Existing DOE data are a treasure-trove of untapped data. It can be further analyzed with the TRA computational software to obtain the many benefits outlined in this chapter. In this context, the new technology is “one large step beyond DOE.”

Table 1 shows a table of five setups (runs) with five parts manufactured for each setup. This creates a sample of 25 parts. The relevant dimensions are measured for each part. The author recommends that a minimum of 25–30 data points (sample parts) be used for the correlation study.

If the part is a new part or there is no existing DOE, then data must be generated for the correlation study. In this instance, variation is induced into the manufacturing process, and parts are made and measured. A noncritical dimension may be the best statistical predictor of the critical dimensions. Consequently, noncritical dimensions can be included in the correlation study. The usual practice to induce variation has been to select the three process parameters that have the greatest impact on part dimensionality and/or performance. These three parameters are then varied in accordance with an experiment designed to meet the user’s requirements.

The five setups in Table 1 are a DOE L_4 with a nominal (center) point. Once the entire system of dimensions is reduced to a single predictor dimension, this facilitates, if one so chooses, learning the relationships between the predictor and the process settings. This enables getting useful results from a DOE program. Alternatively, a three-factor, two-level, full factorial with

Table 1 Inducing Variation to Make 25 Sample Parts

Sample Part No.	Process Variable		
	Pressure	Speed	Time
1 a,b,c,d,e	Low	Low	High
2 a,b,c,d,e	High	Low	Low
3 a,b,c,d,e	Low	High	Low
4 a,b,c,d,e	High	High	High
5 a,b,c,d,e	Nominal	Nominal	Nominal

a centerpoint and three repetitions or replications per run can used.* This design generates 27 data points (parts) and gives all main effects, all interactions, and an indication of linearity if a follow-on DOE study is done. Additional alternative data gathering designs can be used depending on the circumstances. It is not necessary, at all, to do a DOE follow-on study[†] to learn the relationships shown in Fig. 13. The results from a follow-on DOE study are a by-product, albeit a very useful by-product, from the correlation study.[‡]

Table 2 shows the part characteristic data when formatted into a TRA input table. The input table comprises the vehicle for inputting the data into the TRA computational software.

Selecting the Predictor

Table 3 shows the rankings table generated by the TRA computational software. The software ranks all part characteristics from statistically best to statistically worst “Predictor Characteristic” or “Predictor Dimension” using a proprietary algorithm. A substantial number of computations, depending on the number of part characteristics—typically 2000–50,000—are involved. The predictor dimension can be a critical dimension on the engineering drawing or not. In practice, more than one predictor dimension may be selected. In this case, one predictor is selected for each data subset.

Predictive statistical capabilities are based on a comparison of correlation coefficients between all possible combinations of part characteristics. Figures 15 and 16 illustrate possible part characteristic combinations for a single-unit and a multiunit process, respectively. The user has the option of overriding the software to select the second or third best statistical predictor. This alternative is provided in the event the statistically best predictor is difficult, unreliable, or noneconomical to measure. In the example shown in Table 3, variable 1 in data column 1 was the statistically best predictor. However, the user elected to use variable 5 as the predictor because variable 1 required cutting the part open to access it.

Generating the Correlation Charts

Figure 17 shows conceptually how a correlation chart is created. Each data point represents one part. One correlation chart is generated for each predicted dimension. For a linear data set, a linear regression line is fitted through the data set using a least-squares curve-fitting technique. Nonlinear data sets require nonlinear regression lines.[§]

Adding Design Information to the Correlation Charts

Figure 18 shows how design information is added to the correlation charts. There are only two dimensions for each correlation chart. One is the predictor dimension; the other is the critical (predicted) dimension. The critical dimension has a target value (C-TARGET) and the predictor dimension has a target value (P-TARGET). The intersection of these two values is defined as the target intersection.

Figure 18 adds variable names and structure to the correlation chart. The horizontal dotted lines are the upper (USLc) and lower (LSLc) specification limits for the critical dimension. The vertical dotted lines are the upper (USLp) and lower (LSLp) specification limits for the predictor dimension.

* Some users start (run 0) and end (run 9) with a centerpoint run. This gives an indication of process drift during the course of the study.

† The author wishes to stress that we do not have to use DOE to get useful results. In this context, any information we learn about main effects or interactions is “free” or “bonus” information that is a *by-product* of the correlation study.

‡ The potential for having a follow-on DOE study is why a TRA correlation study is occasionally mistaken for a DOE study.

§ Surprisingly, at least on first examination, all data sets examined to date exhibit linear relationships.

Table 2 Measure and Record Dimensions

Run	Cav1 .150	Cav1 .358 gate 90°	Cav1 .455 gate 90°	Cav1 .318	Cav1 .540	Cav1 .478	Cav1 .150	Cav2 .358 gate 90°	Cav2 .455 gate 90°	Cav2 .318	Cav2 .455 gate 90°	Cav2 .540	Cav2 .478
1a	0.1490	0.3575	0.3540	0.4520	0.3165	0.4795	0.1490	0.3575	0.4565	0.3150	0.4525	0.5400	0.4800
1b	0.1490	0.3575	0.3540	0.4520	0.3165	0.4795	0.1490	0.3575	0.4565	0.3150	0.4525	0.5400	0.4800
1c	0.1490	0.3580	0.3540	0.4520	0.3165	0.4795	0.1490	0.3575	0.4565	0.3150	0.4525	0.5405	0.4800
1d	0.1490	0.3575	0.3540	0.4520	0.3165	0.4790	0.1490	0.3575	0.4565	0.3150	0.4525	0.5400	0.4800
1e	0.1490	0.3575	0.3540	0.4520	0.3165	0.4790	0.1490	0.3570	0.4565	0.3150	0.4525	0.5400	0.4800
2a	0.1485	0.3575	0.3540	0.4520	0.3160	0.4795	0.1485	0.3575	0.4560	0.3150	0.4520	0.5400	0.4795
2b	0.1485	0.3575	0.3540	0.4520	0.3165	0.4790	0.1485	0.3575	0.4565	0.3150	0.4520	0.5400	0.4795
2c	0.1485	0.3575	0.3540	0.4520	0.3160	0.4795	0.1485	0.3575	0.4565	0.3150	0.4520	0.5400	0.4795
2d	0.1485	0.3575	0.3540	0.4520	0.3160	0.4790	0.1485	0.3575	0.4560	0.3150	0.4520	0.5395	0.4795
2e	0.1485	0.3575	0.3540	0.4520	0.3165	0.4795	0.1485	0.3575	0.4560	0.3150	0.4520	0.5395	0.4795
3a	0.1490	0.3575	0.3530	0.4555	0.3160	0.4785	0.1485	0.3570	0.4555	0.3150	0.4525	0.5390	0.4790
3b	0.1490	0.3575	0.3530	0.4555	0.3160	0.4785	0.1485	0.3570	0.4555	0.3150	0.4520	0.5390	0.4790
3c	0.1485	0.3570	0.3530	0.4555	0.3160	0.4785	0.1485	0.3570	0.4555	0.3150	0.4520	0.5390	0.4790
3d	0.1490	0.3570	0.3530	0.4560	0.3160	0.4785	0.1485	0.3570	0.4555	0.3145	0.4520	0.5390	0.4790
3e	0.1490	0.3570	0.3530	0.4560	0.3160	0.4785	0.1485	0.3570	0.4555	0.3145	0.4520	0.5390	0.4785
4a	0.1490	0.3575	0.3530	0.4565	0.3160	0.4785	0.1485	0.3570	0.4565	0.3145	0.4520	0.5390	0.4785
4b	0.1485	0.3575	0.3530	0.4560	0.3160	0.4785	0.1485	0.3570	0.4560	0.3145	0.4520	0.5390	0.4785
4c	0.1490	0.3575	0.3535	0.4560	0.3160	0.4785	0.1485	0.3570	0.4560	0.3145	0.4520	0.5390	0.4790
4d	0.1485	0.3570	0.3530	0.4560	0.3160	0.4785	0.1485	0.3570	0.4560	0.3145	0.4520	0.5390	0.4790
4e	0.1490	0.3570	0.3530	0.4560	0.3160	0.4785	0.1485	0.3570	0.4560	0.3145	0.4520	0.5390	0.4790
5a	0.1485	0.3570	0.3530	0.4560	0.3155	0.4780	0.1485	0.3565	0.4555	0.3140	0.4520	0.5390	0.4780
5b	0.1485	0.3575	0.3530	0.4560	0.3155	0.4780	0.1485	0.3570	0.4560	0.3140	0.4520	0.5390	0.4780
5c	0.1485	0.3575	0.3530	0.4560	0.3155	0.4780	0.1485	0.3570	0.4560	0.3140	0.4520	0.5390	0.4780
5d	0.1485	0.3575	0.3530	0.4560	0.3155	0.4780	0.1485	0.3570	0.4565	0.3140	0.4520	0.5390	0.4780
5e	0.1485	0.3570	0.3530	0.4555	0.3155	0.4780	0.1485	0.3565	0.4555	0.3140	0.4520	0.5390	0.4780

Table 3 Rankings Table Specifies Predictor Dimension

Unranked			Ranked		Best Predictor Data Column No.	User Predictor Data Column No.
Col. No.	Variable	Metric	Variable	Metric		
1	Var1	98.0	Var1	98.0	1	5
2	Var2	97.8	Var5	97.9		
3	Var3	97.2	Var2	97.8		
4	Var4	97.7	Var7	97.8		
5	Var5	97.9	Var4	97.7		
6	Var6	97.5	Var6	97.5		
7	Var7	97.8	Var3	97.2		
8	Var8	96.5	Var8	96.5		
9	Var9	95.7	Var10	96.4		
10	Var10	96.4	Var9	95.7		

Best Predictor Variable
Var1

User Predictor Variable
Var5

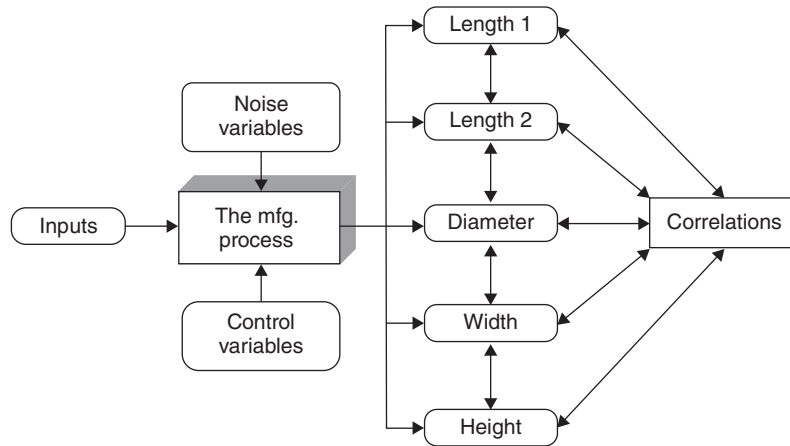


Figure 15 Possible correlations for a single-unit process.

Region of Conformance

The region of conformance is defined as the area bounded by and within the four specification limits in Fig. 19. Good parts lie within the region of conformance. Bad parts lie outside the region of conformance. In practice, the boundaries of the region of conformance are, as a matter of convenience, referred to as the “specification box” or “spec. box.” The regression line can have four basic orientations relative to the region of conformance. Each of the four basic orientations is defined as a condition.

Five Relationship Conditions

Condition 1—Robust Critical. Figure 20 illustrates condition 1, which is defined as “robust.”* The regression line is relatively flat and the correlation coefficient is close to zero. The critical dimension will be within specification limits regardless of whether or not the predictor

* The robust relationship is a subset of a nonconstraining relationship (condition 2).

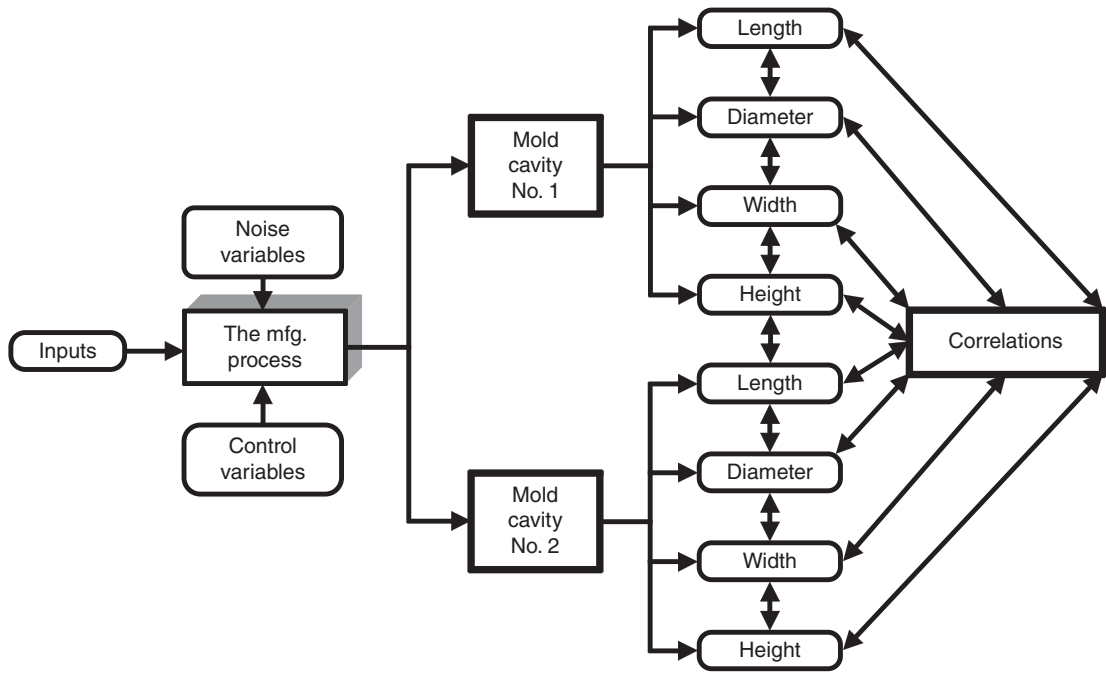


Figure 16 Possible correlations for a multiunit process.

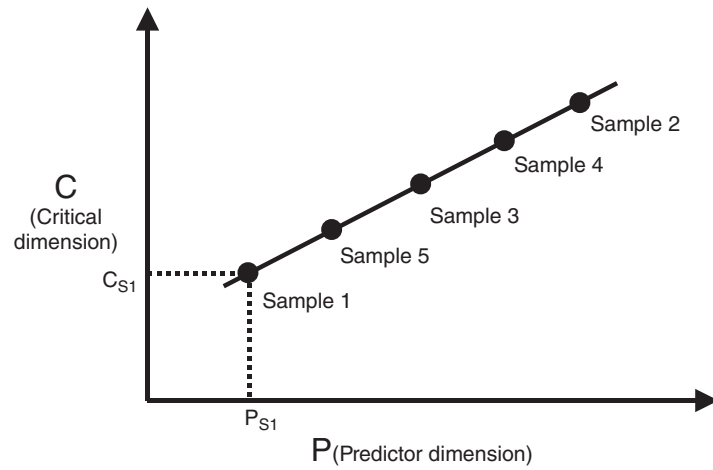


Figure 17 Generating a correlation chart.

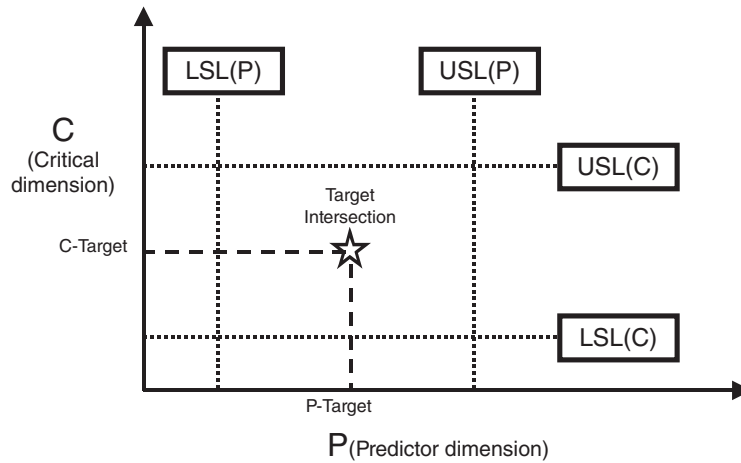


Figure 18 Specification limits and target intersection: one part with two dimensions.

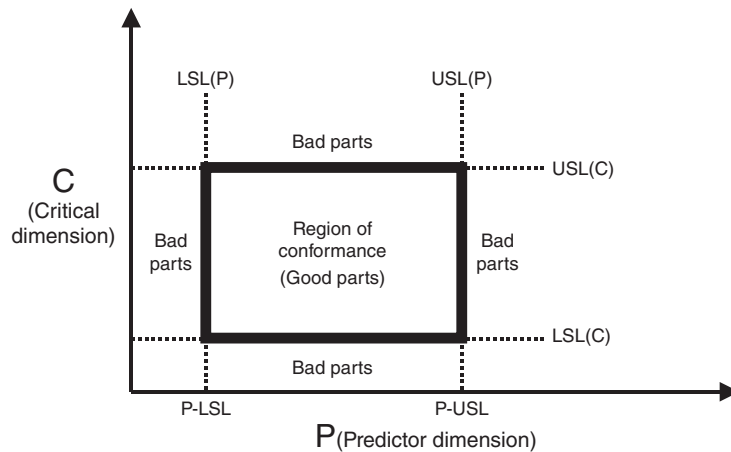


Figure 19 Region of conformance.

dimension is within specification limits.* When the predictor dimension is within specification limits, then the paired dimension values fall within the region of conformance, the part will be a good part. The robust critical dimension never needs to be measured.

Condition 2—Nonconstraining Critical. Figure 21 illustrates condition 2, which is defined as “nonconstraining.” When the predictor dimension is within specification limits, then both the critical and predicted dimensions will be within specification limits. Therefore, when the predictor dimension is within its specification limits, the operating point will be in the region of conformance and the part will be a good part. The nonconstraining critical dimension never has to be measured.

* The correlation coefficient is 0.0 for a zero slope regression line. In this case, the regression relationship is usually considered to be of no practical use. However, TRA uses the zero slope regression line by recognizing that the predicted dimension is insensitive to changes in process settings.

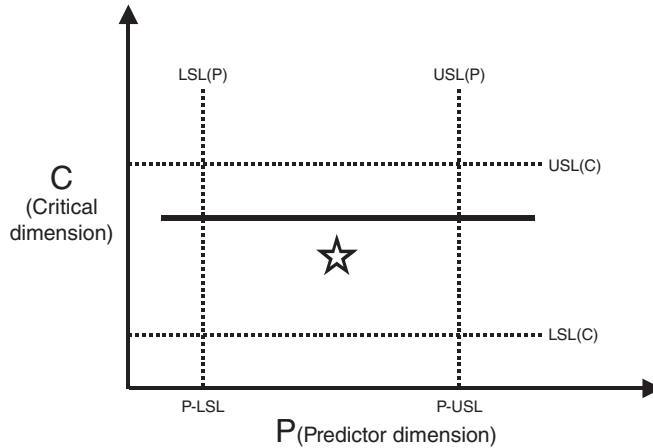


Figure 20 Robust critical dimension (condition 1).

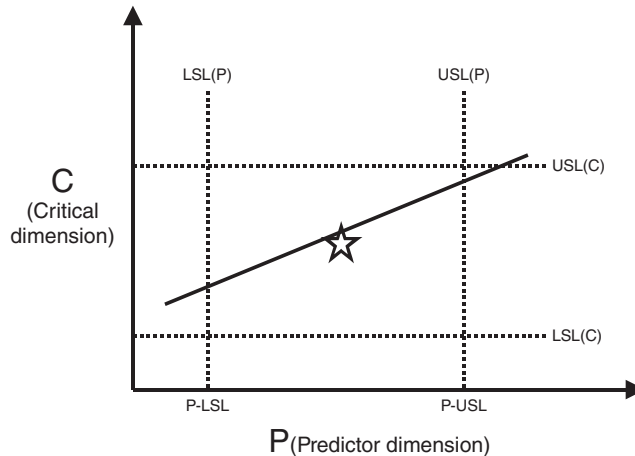


Figure 21 Nonconstraining critical dimension (condition 2).

Condition 3—Constraining Critical. Figure 22 illustrates condition 3, which is defined as “constraining.” When the predictor dimension is between P_{min} and P_{max} , then the constraining critical dimension will be within its specification limits. When this is the case, the constraining critical dimension never has to be measured.

For a regression line with a positive slope, P_{min} is located at the intersection of the regression line and the lower specification limit of the critical dimension. For a regression line with a negative slope, P_{min} is located at the intersection of the regression line and the upper specification limit of the critical dimension.

For a regression line with a positive slope, P_{max} is located at the intersection of the regression line and the upper specification limit of the critical dimension. For a regression line with a negative slope, P_{max} is located at the intersection of the regression line and the lower specification limit of the critical dimension.

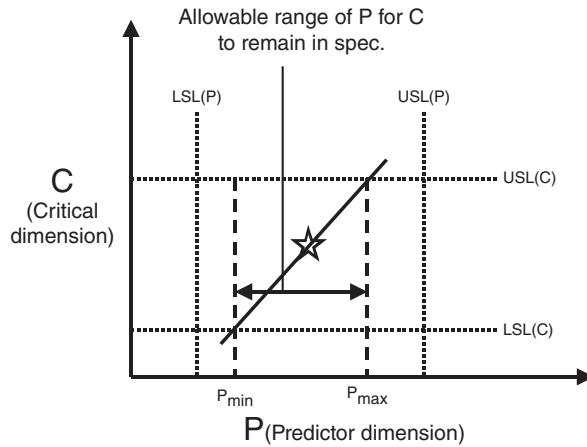


Figure 22 Constraining critical dimension (condition 3).

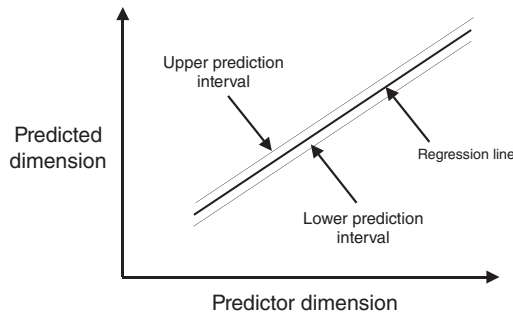


Figure 23 Dispersion of data points (primarily measurement variability).

The preceding discussion can be summarized as follows:

When the predictor is between the greater of LSL_p and P_{min} and the lesser of USL_p and P_{max} , the constraining critical dimension never needs to be measured.

IMPERFECT CORRELATION. In the real manufacturing world, perfect correlation seldom occurs. Figure 23 illustrates imperfect correlation conceptually. Measurement error is one of the main contributors to imperfect correlation. The condition of imperfect correlation is resolved by bounding the regression lines in Figs. 20–22 with upper and lower (three-sigma) prediction intervals shown in Fig. 23.

Figures 24 and 25 are examples of real-world data showing imperfect correlation. The critical dimension is robust when:

- The slope of the regression line is relatively flat.
- The upper and lower prediction intervals are within the region of conformance when the predictor is in between its upper and lower specification limits.

The robust critical dimension still never needs to be measured even though there is imperfect correlation. The critical dimension is nonconstraining when:

- The upper and lower prediction intervals are within the region of conformance when the predictor is in between its upper and lower specification limits.

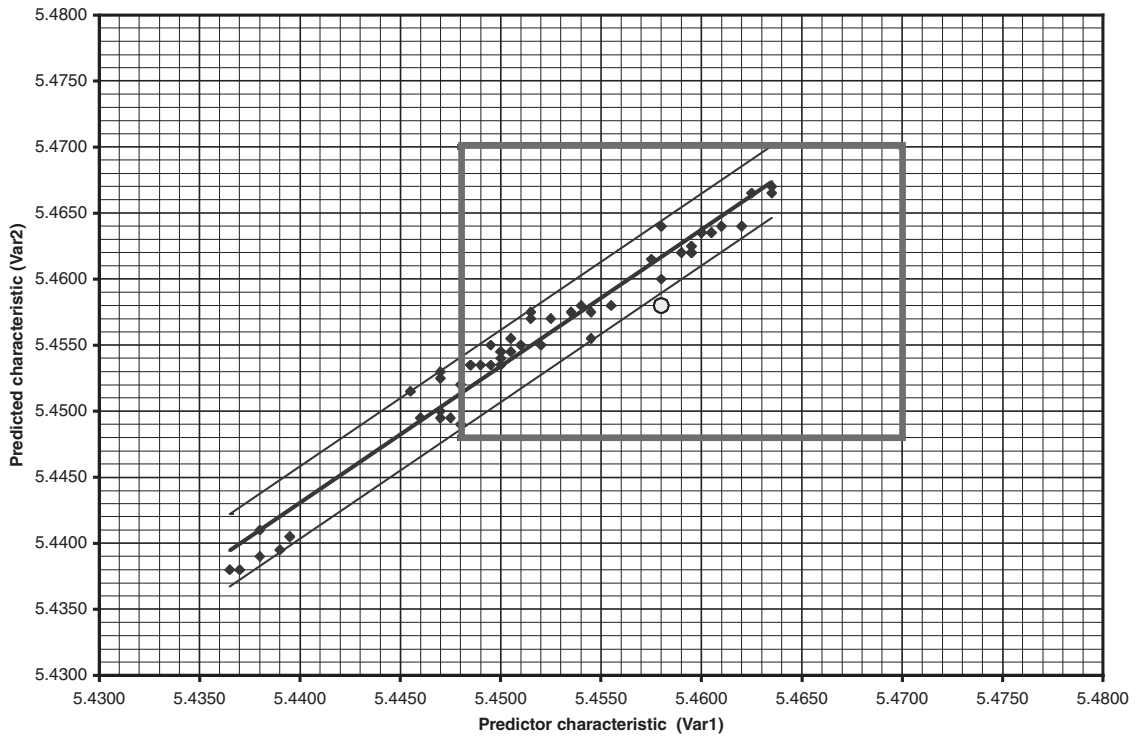


Figure 24 Constraining relationship example.

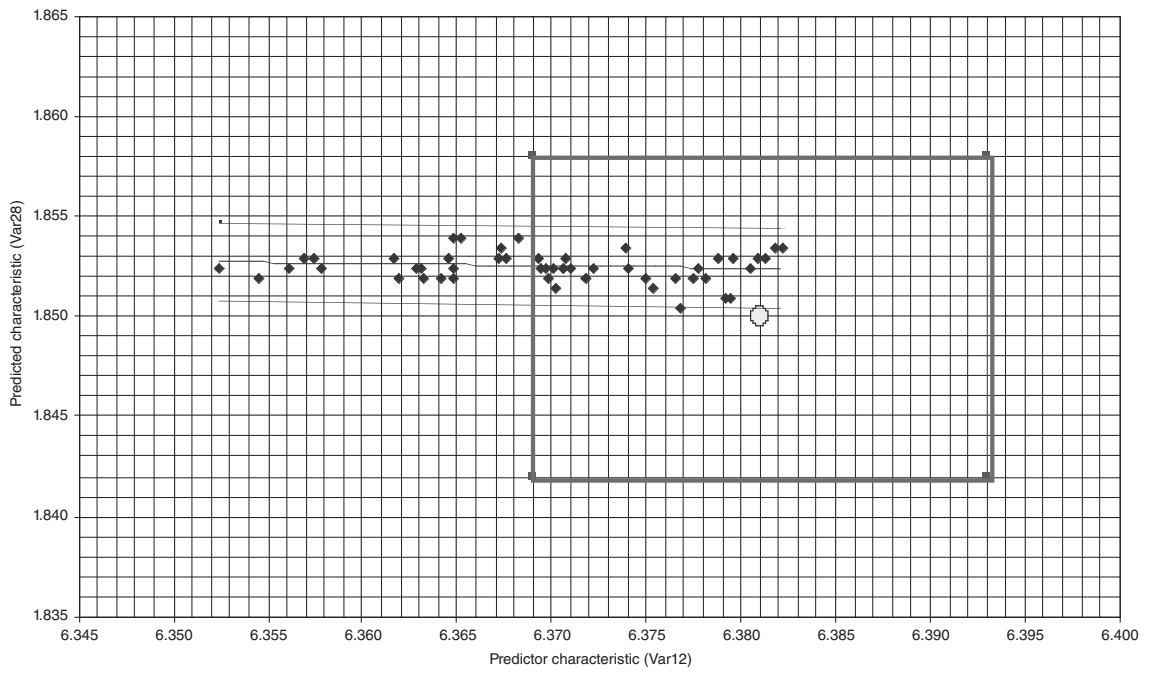


Figure 25 Robust dimension example.

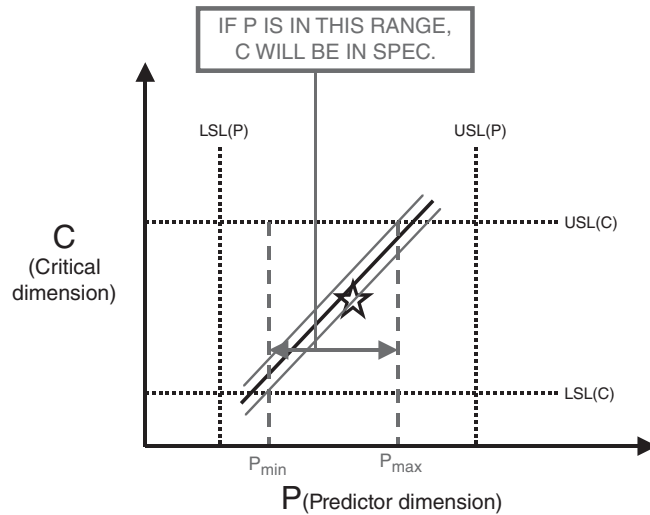


Figure 26 Upper and lower constraints on predictor.

The nonconstraining critical dimension still never needs to be measured even though there is imperfect correlation.

Figure 26 illustrates a constraining condition with imperfect correlation. When the predictor dimension is between P_{min} and P_{max} , then the constraining critical dimension will be within its specification limits. When this is the case, the constraining critical dimension never has to be measured.

For a regression line with a positive slope:

- P_{min} is located at the intersection of the lower prediction interval and the lower specification limit of the critical dimension
- P_{max} is located at the intersection of the upper prediction interval and the upper specification of the critical dimension.

For a regression line with a negative slope:

- P_{min} is located at the intersection of the lower prediction interval and the upper specification limit of the critical dimension.
- P_{max} is located at the intersection of the upper prediction interval and the lower specification limit of the critical dimension.

OPERATING LIMITS. Figure 27 illustrates a part that has three dimensions—two critical dimensions ($C1$ and $C2$) and the predictor dimension (P). With Fig. 22 as a reference, both $C1$ and $C2$ in Fig. 27 are constraining critical dimensions. For the part to be a good part, the predictor must be in between the most constraining (largest) of the P_{min} 's, which is defined as P_{min}^* , and the most constraining (smallest) of the P_{max} 's, which is defined as P_{max}^* .

OPERATING RANGE. The operating range is the distance between P_{min}^* and P_{max}^* . The lower operating limit is the greater of LSL_p or P_{min}^* . The upper operating limit is the lesser of USL_p or P_{max}^* . Figure 28 shows more clearly that the operating range consists of the range of values for the predictor where both $C1$ and $C2$ are within specification limits.

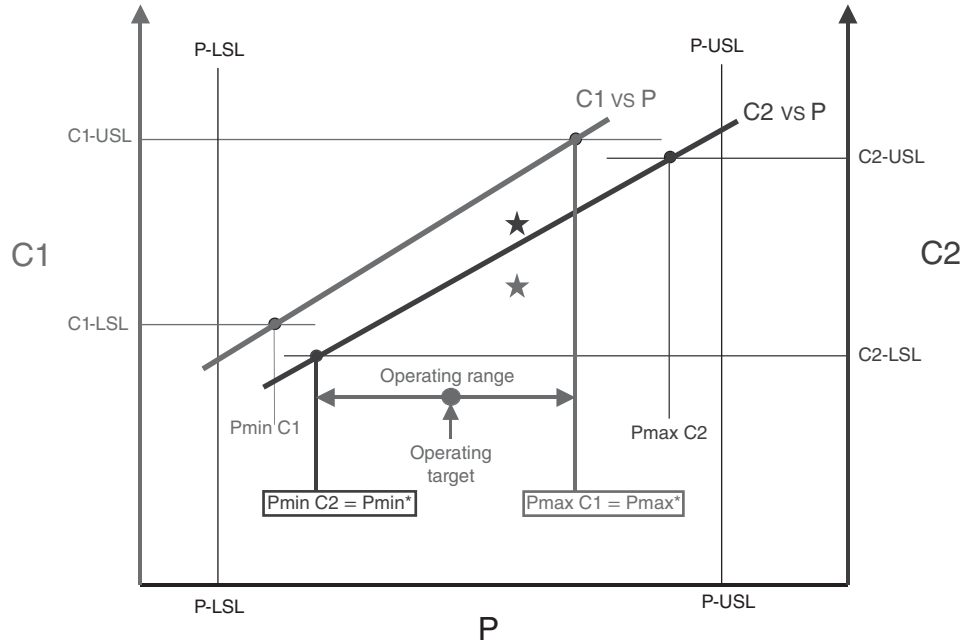


Figure 27 Generalizing to three or more dimensions.

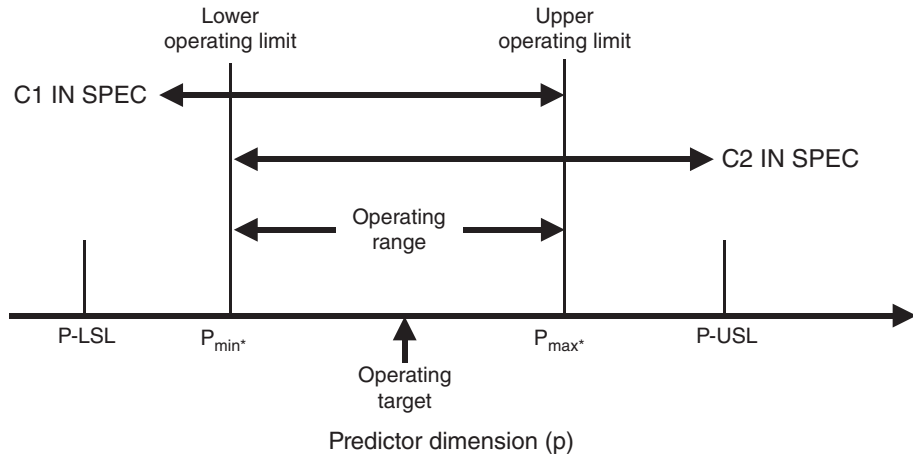


Figure 28 Operating range and operating target.

OPERATING TARGET. Figures 27 and 28 also identify the location of the operating target. For symmetrical process output, the operating target is located at the center of the operating range. For nonsymmetrical process output, the operating target is best selected as the point at which there is equal area in each of the tails of the process output distribution outside of the operating range. Alternative schemes can be used to determine the operating target.

Table 4 Constraint Table

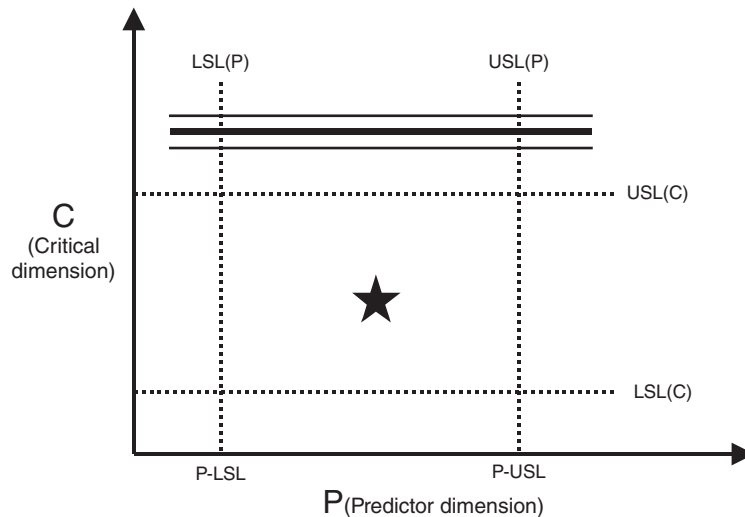
Variable Name	Data Column	Pmin	Pmax	Lower Op. Limit	Upper Op. Limit
Var1	1	5.4538	5.4660	5.4617	5.4647
Var2	2	5.4512	5.4647	9	2
Var3	3	5.4531	5.4660		
Var4	4	5.4497	5.4650		
Var5	5	Predictor	Predictor	Operating Range 0.0030	
Var6	6	5.4511	5.4660		
Var7	7	5.4512	5.4660		
Var8	8	5.4565	5.4658		
Var9	9	5.4617	5.4660	Operating Target 5.4632	5.4658
Var10	10	5.4584	5.4660		

Data is Constrained

CONSTRAINT TABLE. The constraint table shown in Table 4 is a continuation of the previous example shown in Table 3. The constraint table is generated by the TRA computational software. Here are the key elements from the constraint table:

- Variable 5 is the predictor dimension.
- The lower operating limit is 5.4617 in. and is created by variable 9.
- The upper operating limit is 5.4647 in. and is created by variable 2.
- The operating range is 0.0030 in.
- The operating target is 5.4632 in.

Condition 4—Defects—Fixes Needed. Figure 29 illustrates condition 4, which is defined as “defects.” Regardless of the value of the predictor dimension in Fig. 29, the critical dimension will be outside of specification limits and defects will be produced. Similarly, regardless of the

**Figure 29** A defect condition requires change (condition 4).

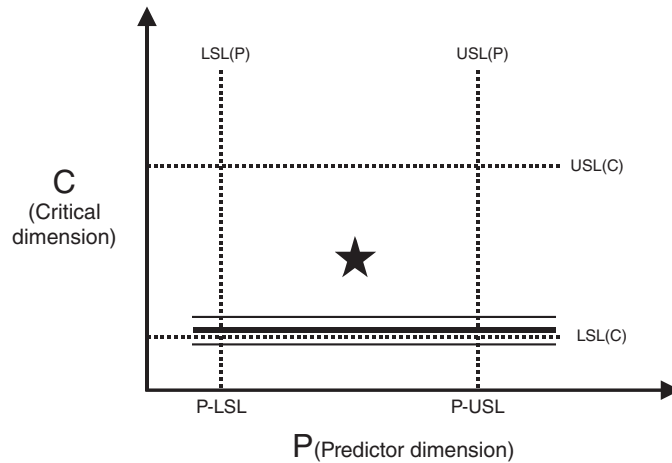


Figure 30 A defect condition requires change (condition 4A).

value of the predictor dimension in Fig. 30, a portion of the critical dimensions are likely to be out of specification and defects are likely to be produced. Figures 29 and 30 can be summarized as follows:

When at least one prediction interval lies outside the conformance area, defective parts are likely to be produced and the defect critical dimension never needs to be measured. Instead, other action must be taken.

If the predicted dimension is

1. robust,
2. nonconstraining, or
3. constraining and the predictor dimension is measured,

then the predicted dimension does not have to be measured and no further action is required. If none of the above three conditions exist, then action must be taken to produce good parts:

Condition 5—Cannot Produce Parts at Design Target. Figure 31 illustrates a situation where it is not possible to produce parts at the target intersection, meaning that it is not possible to produce a part where both the critical and predictor dimensions are at their design target values. This is because the target intersection does not lie within the prediction intervals that bound the data scatter around the regression line.

Fixes

Fix 1—Eliminate Defects by Changing a Design Tolerance. Figure 32 illustrates fix 1. The critical dimension design tolerances are increased to the point where the specification limits encompass the upper and lower prediction intervals. In this instance, the critical dimension's upper (+) design tolerance is increased. The critical dimension design target is unchanged. Referring back to Fig. 27, we see that the operating range will increase as P_{min}^* moves to the left. P_{min}^* is moved to the left by moving P_{min-C2} to the left. P_{min-C2} is moved to the left by reducing (lowering) the lower specification limit for C2 (C2-LSL). The lower specification limit for C2 is reduced by increasing the value of the lower (–) tolerance on C2. At some point, as P_{min-C2} moves to the left, P_{min-C1} will also become a constraining critical dimension.

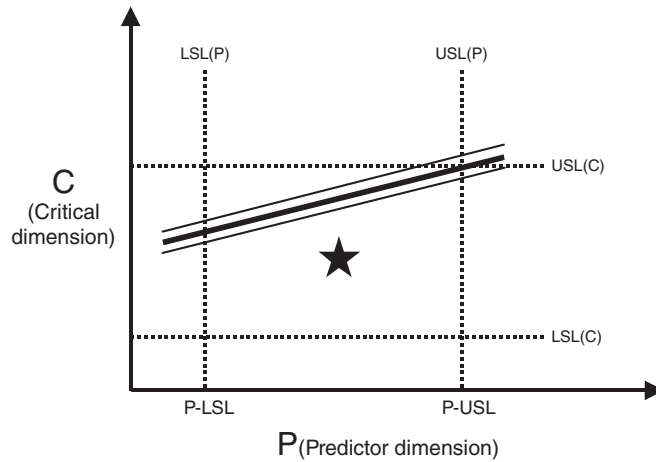


Figure 31 Cannot produce parts at target (condition 5).

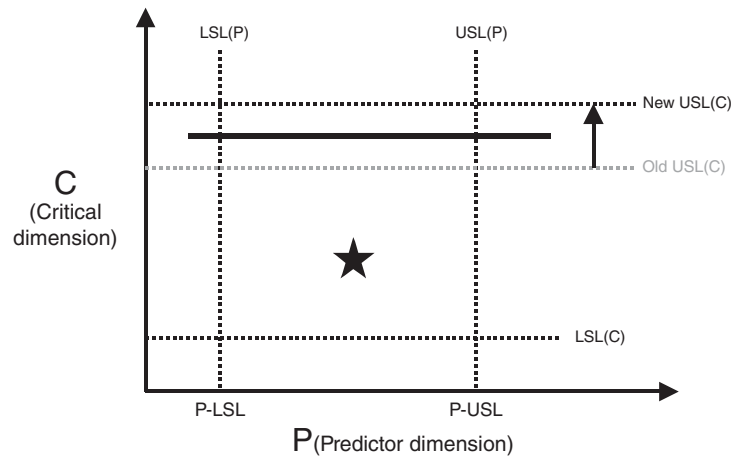


Figure 32 Fix 1: Expand the specification box (change a specification limit, same target).

To continue to increase the operating range beyond this point, the lower tolerances must be relaxed on both C1 and C2.

The same logic applies when there are more than the two predicted dimensions, as illustrated in Fig. 30. Whenever a new constraining dimension is encountered as the operating range is increased, the tolerance on that new constraining dimension will also have to be relaxed. For this reason, a successively larger number of tolerances will have to be relaxed to achieve successively larger increases in the operating range (producibility window). The Pmin tolerance relaxation table is a triangular table for this reason. Table 5 illustrates a sample Pmin tolerance relaxation table. There is a similar table for Pmax. The Pmin and Pmax tolerance relaxation tables provide the design engineer with

- A prioritized ranking of the order in which design tolerances must be relaxed
- The required increase in design tolerances required to achieve any specified increase in the producibility window

Table 5 Pmin Tolerance Relaxation Table

Limiting Variable	Old Pmin	New Pmin	Individ. Gain	Cumul. Gain	New Tolerances										
					Var9	Var10	Var8	Var1	Var3	Var2	Var7	Var6	Var4	Var5	
Var9	5.4617	5.4584	0.0033	0.0033	6.3655										
Var10	5.4584	5.4565	0.0019	0.0052	6.3635	6.3668									
Var8	5.4565	5.4538	0.0027	0.0079	6.3607	6.3637	5.4451								
Var1	5.4538	5.4531	0.0007	0.0087	6.3599	6.3628	5.4444	5.4472							
Var3	5.4531	5.4512	0.0018	0.0105	6.3580	6.3607	5.4424	5.4453	5.4461						
Var2	5.4512	5.4512	0.0000	0.0105	6.3580	6.3606	5.4424	5.4453	5.4460	5.4479					
Var7	5.4512	5.4511	0.0001	0.0107	6.3578	6.3605	5.4423	5.4451	5.4459	5.4478	5.4479				
Var6	5.4511	5.4497	0.0014	0.0121	6.3564	6.3589	5.4408	5.4437	5.4444	5.4463	5.4464	5.4467			
Var4	5.4497	5.4480	0.0017	0.0137	6.3546	6.3570	5.4390	5.4419	5.4427	5.4444	5.4447	5.4452	5.4463		
Var5	5.4480	5.4480	0.0000	0.0137	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

First Pmin Tolerance Relaxation Doubles Operating Range in This Case

Var9 original lower spec. limit = 6.369"

Var8 original lower spec. limit = 5.448"

There is a similar table for the Pmax's.

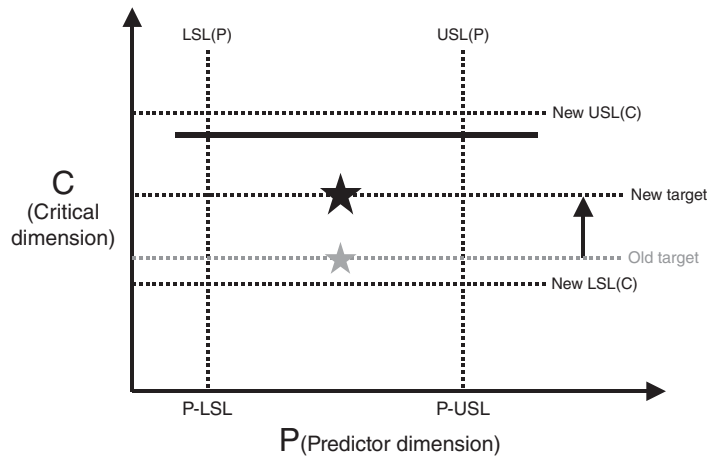


Figure 33 Fix 2: shift the specification box (change the target, same specification limits).

The Pmin and Pmax tolerance relaxation tables eliminate design engineer dependency on operator process settings when determining tolerance relaxations.

Fix 2—Eliminate Defects by Changing the Design Target. Figure 33 illustrates fix 2. The critical dimension target value is changed. This moves the entire specification box in the desired direction. In this instance, the design target is increased, which shifts the specification box upward to the point where the specification box encompasses the upper and lower prediction intervals. The design tolerances are unchanged. The size of the specification box is unchanged.

Design engineers find it practical to change the design tolerance when

- there is no impact on fit or function or
- the target value of a mating dimension can be changed.

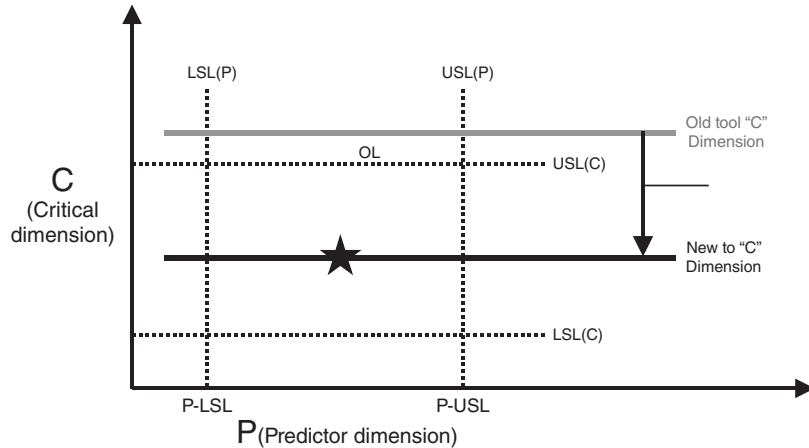


Figure 34 Fix 3: shift the regression line (change the tooling, same target, same specification limits).

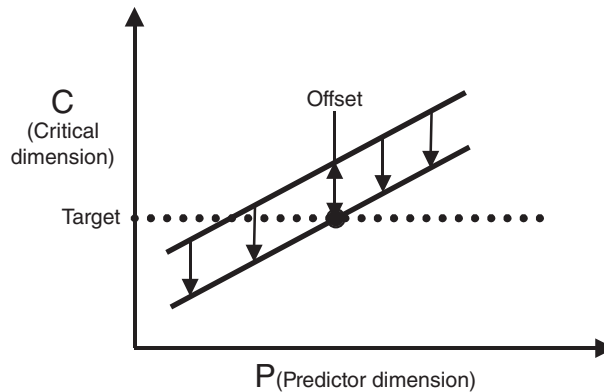


Figure 35 Determine tooling changes independently from operator changes to process settings.

In some cases, tooling engineers find it easier to change the tooling on a mating part. This also is a situation where the design target can be changed.

Fix 3—Eliminate Defects by Shifting the Regression Line. Figure 34 illustrates a shift in a robust regression line. The regression line is shifted by modifying tooling or other preprocess dimensions so that the regression line passes through the target intersection. The offset is defined in Fig. 34 as the vertical (or other) distance between the regression line and the target intersection. Figure 35 illustrates a shift for a nonzero slope regression line.

Table 6 shows the offset table. The offset table is generated by the TRA computational software. It specifies, for each predicted dimension, the vertical shift required for the regression line to pass through the target intersection. The offset table eliminates:

- Tooling engineer dependency on operator process settings
- Multiple cycles of tooling, mold, and fixture changes when going from preproduction to production tooling

Table 6 Offset Table Specifies Optimum Changes to Tooling

Variable Name	Vertical Offset
Var1	-0.0026
Var2	0.0010
Var3	-0.0011
Var4	0.0017
Var5	Predictor
Var6	-0.0012
Var7	-0.0001
Var8	-0.0023
Var9	-0.0096
Var10	-0.0072

Fix 4—Eliminate Defects by Improving Measurement Accuracy. Data scatter about the regression line can be reduced by increasing measurement accuracy. The reduced scatter reduces the size of the prediction intervals. This has the potential of moving both prediction intervals inside of the specification box.

Fix 5—Produce Parts at Target. As shown in Fig. 31, condition 5 exists when it is possible to produce good parts, but it is not possible to produce parts where both the critical and predictor dimensions are at the target value. This is because the target intersection lies outside of the prediction intervals. The target intersection can be brought within the prediction intervals by changing the design target(s) or shifting the regression line.

2.4 Material Selection

Figure 36 shows the producibility window for three different materials. The larger the producibility window, the easier it is to produce the part. Replacements for obsolete material can be evaluated. Design and manufacturing engineers can evaluate the trade-off between producibility and cost. In the best case, the material with the greatest producibility window has the least cost.

Conflict Reduction

TRA reduces conflict and increases collaboration and communication between

- Design engineers
- Tooling engineers

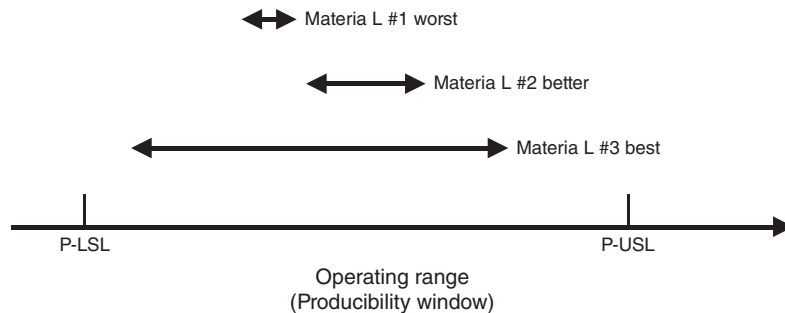


Figure 36 Different materials have different producibility windows.

- Process engineers
- Quality engineers
- Customers and suppliers

Improved Decision Making

TRA greatly improves decision making by:

- Eliminating the effect of process complexities
- Visual presentation of results
- Using a scientific method instead of trial and error, iteration, and guesswork
- A systems approach that integrates all of the engineering functions

TRA Process Flowchart

Prior art techniques use trial and error, iteration, and guesswork to transition from a preproduction condition where good parts can be difficult or impossible to produce to a production condition where good parts can be produced. Figure 37 is a process flowchart of TRA. Tooling, design tolerances, design targets, and material selection are optimized in a single step with TRA. Figure 37 also shows the relationship of the TRA computational software to existing measurement, SPC, process capability, and DOE software. TRA does not replace this software. Instead, the TRA computational software is complementary to the two major functions of existing software as follows:

- The TRA computation software greatly reduces the amount of measurement, SPC, and Cpk analysis that must be performed by the existing software.

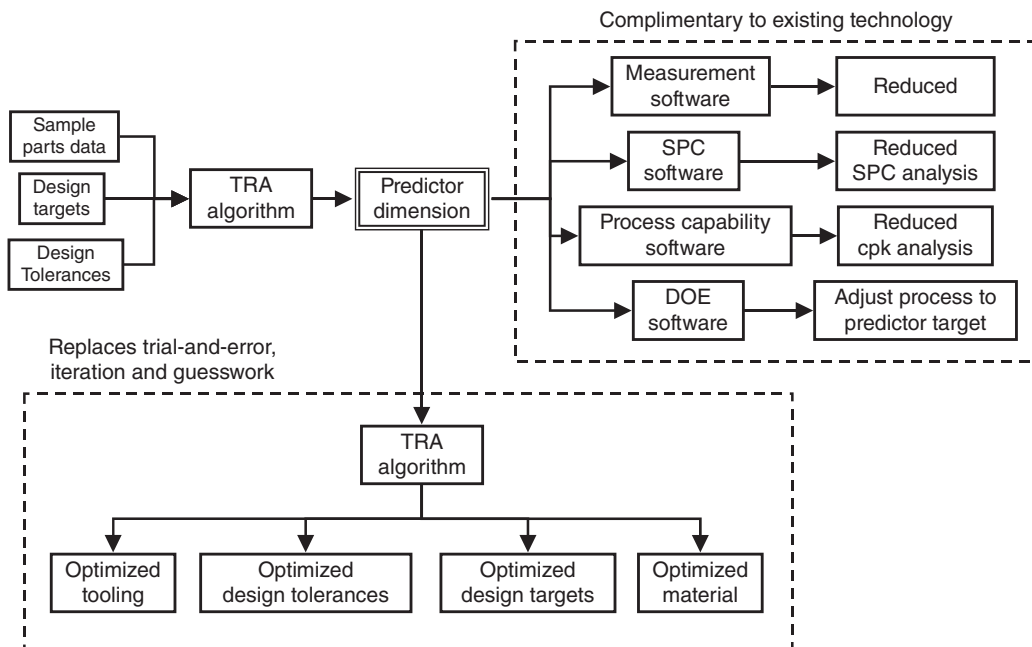


Figure 37 TRA is complimentary to existing technology.

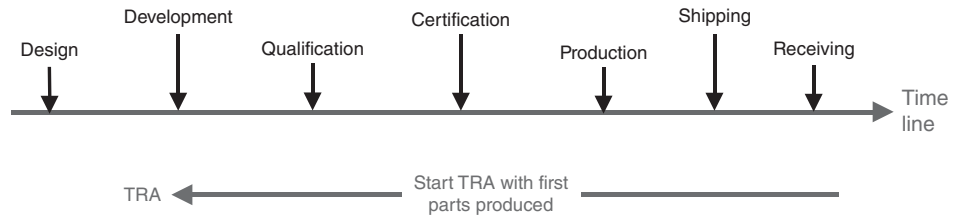


Figure 38 TRA gives cost savings at all stages of development and production.

- The TRA computational software enables DOE software to be used in situations where there are complexities that normally prevent the use of DOE to understand the relationship between causes (process settings) and effects (part characteristics).

When Should TRA Be Used?

Figure 38 shows that the earlier the TRA is used, the quicker the time savings, the financial savings, and the quality improvement will be realized. TRA yields the greatest benefits if it is used when parts are first produced by the process.

Must the Process Be in Control during Production?

The process does *not* have to be in control during production when:

- The process is a single-unit process.
- The process is a dependent multiunit process.
- The out-of-control situation is runs, periodicity, trends, hugging, outside of the three sigma limits, etc.

The process *does* have to be in control during production if the process is an independent multiunit process, which means that the out-of-control situation could affect only one or several of the multiunit processes. Figure 39 shows that there is a substantial savings in measurement and analysis costs in either event.

Simulation Benefits

The computation software also simulates the operation of the manufacturing system used to produce the parts and the measurement system used to measure them. In the simulation mode, the software makes it possible to assess the impact of contemplated changes to design targets, design tolerances, and tooling on manufactured part dimensions without incurring the cost, time, and risk of changing tooling or design parameters and then producing, measuring, and analyzing the new parts.

The left-hand chart in Fig. 40 shows data points generated by a CNC laser cutter.* As can be seen, some of the data points are right on the edge of being bad parts. In this instance, the CNC programming was set to produce parts at the target intersection. The TRA computational software identified biases in the CNC laser cutter. These biases were fed back into the TRA software. The input data were adjusted to simulate the results of the changes that would be made to eliminate the machine/programming biases. The results are shown in the

* These data were generated under manufacturing conditions, meaning that variation was not induced. The data scatter represents the normal process variability.

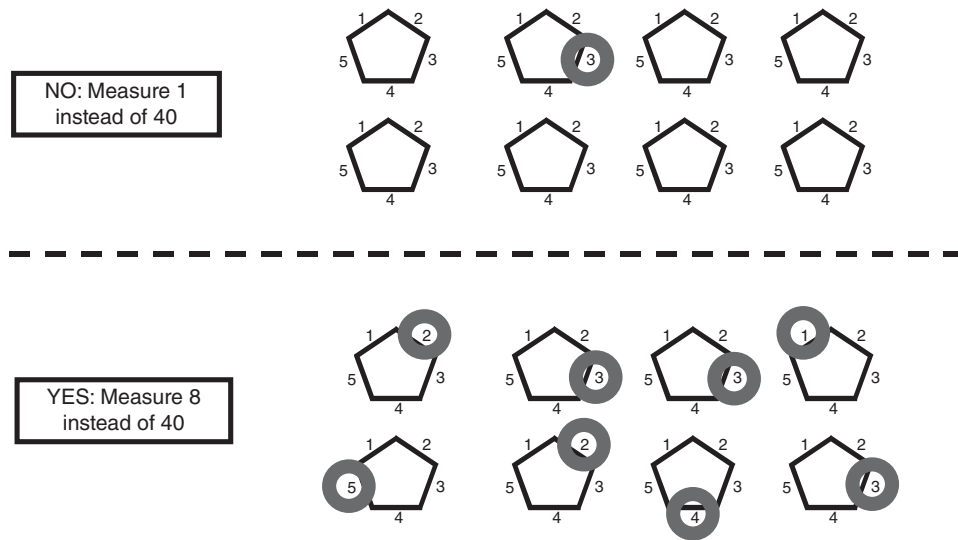


Figure 39 Must the process be in control during production?

right-hand chart in Fig. 40. The data points are now clustered around the target intersection. The TRA computational software has simulated changes to the CNC programming. Changing CNC programming is relatively low cost and risk free. That may not be the case for changing molds, fixtures, and other hard tooling.

3 CONCLUSION

Figures 41–43 show the historical evolution of

- Quality,
- Manufacturability, and
- reduction in time to market

relative to the contributions made by TRA. The algorithms discussed in this chapter are the new models for the manufacture of products in many industries. They provide an immediate increase in profitability and market competitiveness by reducing cost and increasing quality during production and reducing time, cost, and risk during development.

4 IMPLEMENTATION

The TRA algorithms and computational software have received U.S. patent no. 6,687,558. Other patent applications are pending. A demonstration version of the software is available at www.algoryx.com. TRA provides useful insights into manufacturing processes. However, once more than a few part characteristics are involved, the computations and graphical analysis become tedious and labor intensive. The computation software is available from Algoryx, Inc. at the following address: Algoryx, Inc., 750 S. Bundy Drive, Ste. 304, Los Angeles, CA 90049, 310-820-0987, steve@algoryx.com.

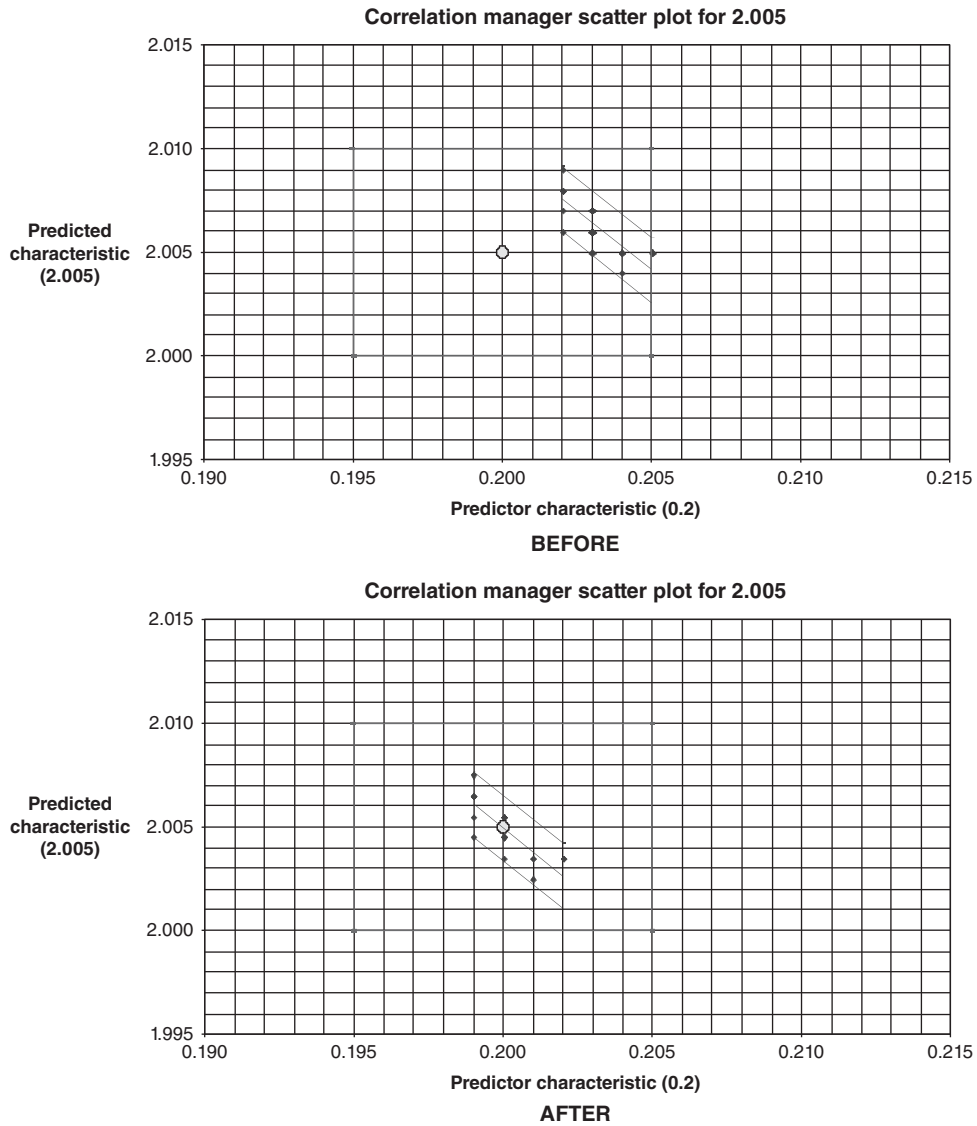


Figure 40 TRA computational software can be used to simulate changes.

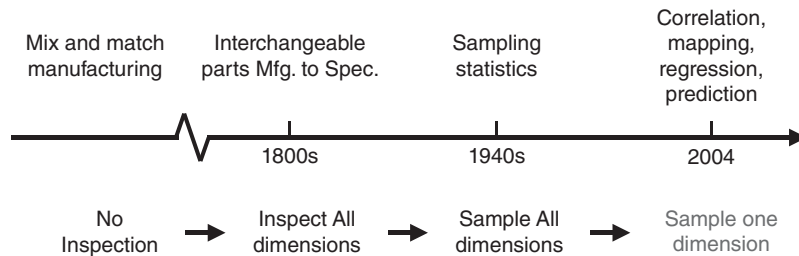


Figure 41 Evolution of quality.

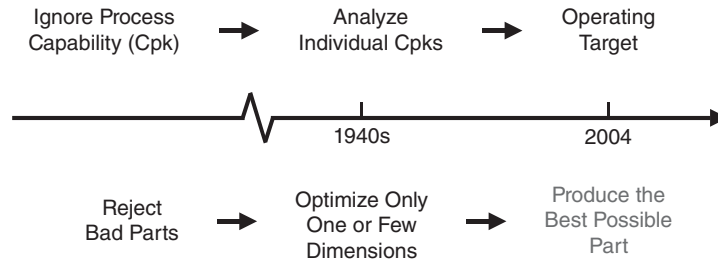


Figure 42 Evolution of manufacturability and product performance.

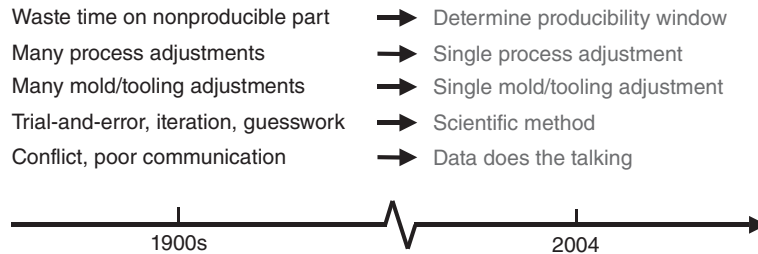


Figure 43 Evolution in reducing time to market.

APPENDIX A: TUSZYNSKI'S PROCESS LAW

1. Every process has causes and effects.
2. Causes can be categorized into (i) slider variables, (ii) shifter variables, and (iii) noise variables.
3. Although the relationships between slider variables and effects may be difficult, economically infeasible, or impossible to predict, the relationships between effects are consistent and predictable.
4. Slider variables can be used to reposition the operating points within relationships.
5. Shifter variables can be used to change the relationships between effects.
6. The system of relationships has one degree of freedom; one effect can be used to predict one or more other effects.
7. The effect best able to predict all other effects is the predictor effect.
8. The relationship between any two effects, including the predictor, is unique.
9. When effects have tolerances that bound their region of conformance, each effect is, relative to the predictor,
 - a. robust,
 - b. nonconstraining,
 - c. constraining, or
 - d. nonconforming.
10. Effects never have to be measured when they are:
 - a. Robust
 - b. Nonconstraining and the predictor is conformal

- c. Constraining and the predictor is within the operating range
- d. Nonconforming
- 11. Process performance is optimal when the predictor is at the operating target.
- 12. Effects can be produced at the target intersection only when the target intersection is within the regression area.
- 13. Nonconforming effects can be made conforming by any combination of:
 - a. Relaxing tolerances
 - b. Changing design targets
 - c. Changing shifter variables
 - d. Constraining slider variables
 - e. Reducing measurement error
- 14. If shifter variables change predictably with time, the number of process cycles, or any other variable, the resultant changes in relationships are predictable.

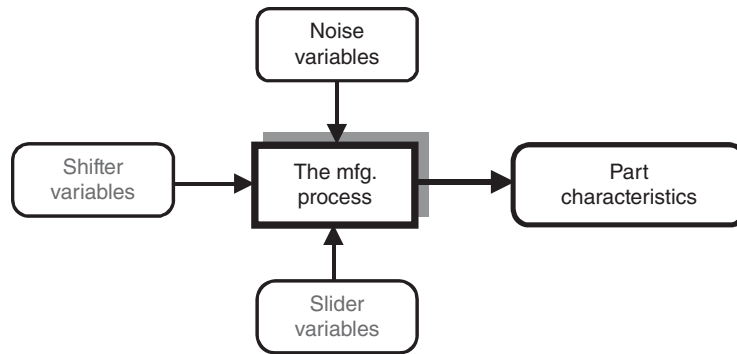


Figure 44 TRA process flow diagram.

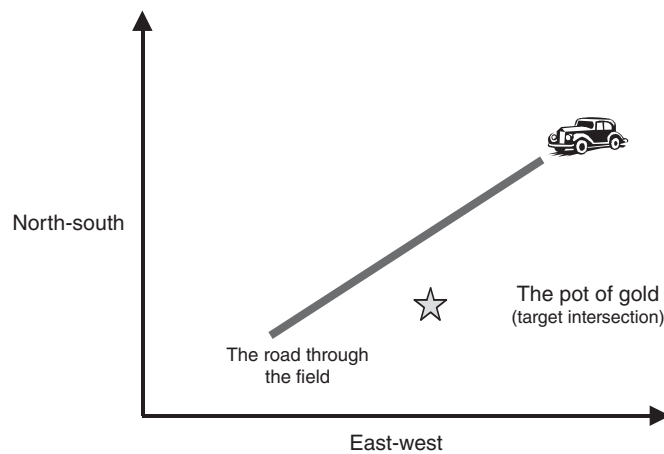


Figure 45 You can't get there from here.

15. Tuszynski's process laws apply to:
- Single-unit processes whether the process is in control or not
 - All units jointly of multiunit-dependent processes whether the process is in control or not
 - All units jointly of multiunit processes when the process is in control
 - Individual units severally of a multiunit process when the processing units are independent and the process is out of control

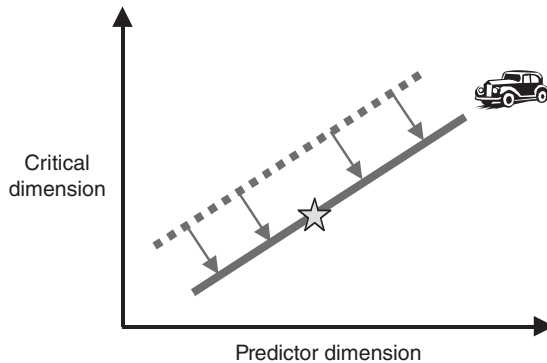


Figure 46 Reroute the road: change “shifter” variables (mold dimensions).

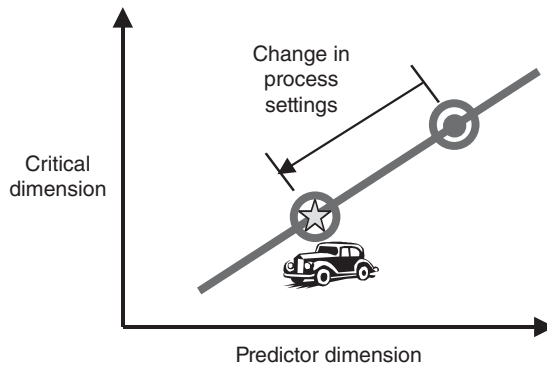


Figure 47 Drive along the road and grab the gold: change “slider” variables (process settings).

APPENDIX B: DEFINITIONS

Cause: A factor that brings about an effect—typically a process setting.

Constraining effects: Constraining effects place upper and lower constraints on the predictor effect. These constraints are constant and are inside the predictor effect tolerances.

Degrees of freedom: The number of parameters (variables) that can be independently varied.

Effect: A result—more specifically, a feature or performance characteristic.

Noise variables: Variables that are uncontrolled because either we do not have the ability to control them or we choose to not control them.

Nonconforming effects: Nonconforming effects are nonconforming when the predictor effect is conformal.

Nonconstraining effects: Nonconstraining effects are conformal when the predictor effect is conformal.

Operating points: Operating points are the points on the regression prediction models (lines).

Operating range: The operating range is the region between the maximum lower constraint and the minimum upper constraint on the predictor effect.

Operating target: The operating target is the point in the operating range that maximizes the percentage of process output in the operating range. For symmetrical processes, the operating target is located at the center of the operating range.

Prediction intervals: Upper and lower boundaries around the regression prediction line that bound the data set—typically at plus/minus three sigma values.

Regression area: The area around the regression model bounded by the prediction intervals.

Relationships: The relationships (correlations) between variables—typically a regression prediction model, which can be linear or nonlinear.

Robust effects: Robust effects are conformal irrespective of the value of the predictor effect.

Shifter variables: Shifter variables change the relationships.

Slider variables: Variables that move the operating points along the relationships. Also called process variables.

Target intersection: The intersection of design target values for effects.

APPENDIX C: NONTECHNICAL STATISTICAL GLOSSARY

Attribute data: Usually refers to a data describing a part characteristic not measurable on a number scale, such as on/off, present/absent, supplier A, B, or C, material type m, n, or o, etc. Some attributes, such as color, can be converted to variable data (a combination of red-blue-green) if it is worth the cost and effort. Also, see *variable data* in this glossary.

Causation: The act or agency that produces an effect. In some manufacturing processes, there is a direct relationship between a process control variable and the process output. In other manufacturing processes, the relationship between process control variables and the process output can be extremely complex. Also, see *relationship vs. causation* in this glossary.

C-LSL: The lower specification limit of a critical dimension.

Coefficient of determination: The square of the correlation coefficient. It represents the percentage of variation in Y that is accounted for by variation in X . If the coefficient of determination is 75% (0.75), then 25% of the variation in Y is caused by a factor or factors other than X . Also, see *correlation coefficient* in this glossary.

Complex interaction: The effect of a change in one process control variable on a part characteristic depends on the level at which two or more different process control variables are set. Also, see *simple interaction* in this glossary.

Constraint table: A table showing, for each critical characteristic, Pmax and Pmin or no constraints (P-USL and/or P-LSL).

Control limits: Limits drawn on a control chart that are a function of the sample size and the variability within the data.

- Correlation coefficient:** A number ranging between +1.0 and -1.0 that represents the degree of relationship between two variables from perfect positive to zero to perfect negative correlation. Plus means a positive slope and minus means a negative slope. Zero or very low correlation can be caused by either a regression line with zero slope or by two variables that exhibit no relationship.
- Critical dimension:** A dimension that the designer feels is critical to the form, fit, or function of the part. Also called a “critical” or “criticals.”
- C-target:** The nominal or target value of the critical dimension set by the design engineer that presumably optimizes form, fit, or function. A point.
- C-USL:** The upper specification limit of a critical dimension.
- Design of experiments (DOE):** A set of techniques to extract useful information from a relatively small number of experimental runs instead of testing all possible combinations of variables. Used to (1) optimize the process that makes a product, (2) make the manufacturing process robust, (3) optimize product performance, and (4) make product performance robust.
- Go/no-go gage:** Gives an accept/reject good part/bad part decision without needing to measure the actual numerical value of the predictor dimension (or any other dimension). This saves time and money.
- Graphical (visual) linear regression:** Drawing a straight line (also called the regression line) through the center of the data points on a scatter plot. The line is visually positioned so that the data points are equally centered above and below the line.
- Graphical prediction intervals:** Typically drawn as two straight lines equidistant from and parallel to the regression line through the outermost data points.
- Hot measurement:** Allows the measurement of a part shortly after it is produced instead of waiting hours or days. Allows production to be stopped or the process to be modified if bad parts are being produced.
- In-control process:** A process where all (99.74%) of the data points are within the control limits and there are no nonnormal patterns such as trends or shifts. See also *out-of-control process* in this glossary.
- Joint operating position:** A point on the regression line (or within the bounds of the prediction intervals).
- Mathematical linear regression:** Drawing a straight line through the center of the data points on a scatter plot. Typically done by computing the slope of the line and one of the intercepts (Y or X).
- Mathematical prediction intervals:** Bounds on the scatter of the data points around the regression line calculated using statistical formulas.
- Natural process limits:** Typically the plus and minus three sigma points on the distribution representing the actual process output.
- Nominal process settings:** Process settings that result in all critical dimensions being within specification limits. Ignores the difference between the average process output(s) and the design target(s) and does not attempt to minimize variation between parts or between cavities. See also *optimum process settings* in this glossary.
- Noncritical dimension:** A dimension that the designer feels is not critical to the form, fit, or function of the part.
- Nonlinearities:** Increasing a process control variable can cause a dimension to initially increase and then decrease.

Nonlinear regression lines: A linear (first-order) regression line is of the form $Y = a + bX$. A second-order regression is of the form $Y = a + bX + cX^2$. Regression lines are seldom fitted higher than third order.

Offset table: A table showing how many units the regression line is above or below the target intersection.

Optimum process settings: Process settings that minimize the difference between the average process output(s) and the design target(s) and minimize variation between parts or between cavities. See also *nominal process settings* in this glossary.

Out-of-control process: A process where data points are outside of the control limits and/or there are nonnormal patterns such as trends or shifts. See also *in-control process* in this glossary.

Part characteristic: Something physical about a part, such as a dimension, weight, color, or hardness. Can be variable or attribute data. Also referred to as a dimension for ease of reading.

P-LSL: The lower specification limit of the predictor dimension.

Pmax.: The point on the X axis—the predictor dimension—determined by the intersection of the upper prediction interval and the upper specification limit of the critical dimension. A point.

Pmax*: The smallest (lowest) value Pmax from the constraint table. The most constraining of the Pmax's. A point.

Pmin*: The point on the X axis—the predictor dimension—determined by the intersection of the lower prediction interval and the lower specification limit of the critical dimension. A point.

Pmin*: The largest (highest) value Pmin from the constraint table. The most constraining of the Pmin's. A point.

Population: The complete set of items. For example, (1) all the parts of the same type produced by a process during a given production run or (2) all the parts of the same type produced by a process during the life of that process.

P-range*: The difference between Pmin* and Pmax*. A distance.

Prediction intervals: Two lines that are equidistant above and below the regression line. The prediction intervals bound a specified percentage of the data points on a scatter diagram. The upper prediction interval is above the regression line. The lower prediction interval is below the regression line. The prediction intervals give the uncertainty range of the predicted variable.

Prediction using regression lines: Given X , Y can be predicted. Given Y , X can be predicted.

Predictor characteristic: A part characteristic selected to predict other part characteristics. Also called a “predictor.”

Predictor dimension: A predictor characteristic that happens to be a dimension. Although called a “predictor dimension” in some of my material, the more accurate description would be predictor characteristic. For example, the predictor dimension could be a weight (which is a characteristic, not a dimension). The use of the word “dimension” is used to improve readability.

Press setting: A control setting on an injection-molding press that controls some element of the press operation. These are typically pressures, temperatures, times, and speeds.

- Process capability study:** A comparison of the distance between the process upper and lower natural limits (plus and minus three sigma) to the distance between the specification limits (TOL). The offset of the average process output from the design target can either be considered (Cpk) or ignored (Cp).
- Process characteristic:** Something physical about the process that makes the part. Typically process characteristics are pressure, temperature, time, speed, chemical concentration, part orientation, etc. Can be variable or attribute data. Some process characteristics are controlled. Others are not (“noise”). Also called process variables or process control variables or process conditions. Press settings are examples of process variables.
- Process control variable:** A process variable that is controlled or set by the operator. A press setting for an injection-molding press. Changing a process control variable typically shifts the joint operating position along the regression line.
- Process input:** Anything that influences the output of a process other than process control or process noise variables. The mold cavity dimensions would be examples of process inputs for an injection-molding process. The preplated dimensions of a part would be examples of process inputs for a plating process. Changing a process input typically shifts the regression line.
- P-target:** The nominal or target value of the predictor dimension set by the design engineer that presumably optimizes form, fit, or function. Equivalent to C-target. A point.
- P-target*:** The value at which it is desired to have the average output of the process. This is typically the center of P-range* for a process with a symmetrical output distribution. A point.
- P-USL:** The upper specification limit of the predictor dimension.
- Range:** The difference between the highest and lowest values in a sample.
- Relationship versus causation:** Scatter plots and correlation coefficients are used to illustrate and determine the degree of relationship between two variables. Scatter plots and correlation coefficients are not used to say that a change in *X* caused a change in *Y*. Causation may or may not be present but must be proved by other means.
- Robust:** To make something robust is to make it resistant to outside influences.
- Sample:** A subset of a population, hopefully taken using valid sampling methods (randomized, unbiased, etc.). When sampling from a production line, samples are gathered in subgroups. The samples within a subgroup should be taken as close together in time (or in order of manufacturing sequence) as possible. The time between subgroup samples is determined by the total number of samples to be taken for a given period of time.
- Scatter plot:** Also known as scatter or correlation or *X-Y* plots. Can be called plots or diagrams or charts. Used to graphically (visually) illustrate the degree of relationship between two variables (*X* and *Y*).
- Sigma:** Shorthand for one standard deviation unit. One sigma unit is a distance. Also known as a “*Z*” unit for a nondimensional normal curve.
- Simple interaction:** The effect of a change in one process control variable on a part characteristic depends on the level at which a different process control variable is set. For example, increasing pressure when temperature is low will increase a dimension, while increasing pressure when temperature is high will decrease that dimension.
- Specification limits:** The upper and lower limits within which a part characteristic must be manufactured in order for the part to adequately meet the requirements of form, fit, and function. These are specified by the design engineer. They are points.

Standard deviation: A measure of the “spread” or “width” of a distribution. A distance. Other measures of the variability of a distribution are the range, average deviation, variance, and semi-interquartile range.

Target: The nominal value of a part characteristic established by the design engineer that optimizes form, fit, or function. A point. Manufacturing first tries to produce good parts (within specification) and then to produce parts that are on target with minimum variation.

Target intersection: The point on a scatter chart determined by the intersection of the *Y* axis (critical) target and the *X* axis (predictor) target.

Tolerance (TOL): The difference between the upper and lower specification limits. A distance.

Variability (VAR): The difference between the upper and lower natural process limits. Typically taken as the distance between the plus and minus three sigma values of the process output. A distance.

Variable data: Refers to data describing a part characteristic measurable on a continuous number scale. Examples are dimensions, weights, hardness, tensile strength, etc.

CHAPTER 15

NONDESTRUCTIVE INSPECTION

Robert L. Crane
Air Force Research Laboratory
Wright Patterson Air Force Base
Dayton, Ohio

Giles Dillingham
Brighton Technologies Group
Cincinnati, Ohio

1 INTRODUCTION	441	5 MAGNETIC PARTICLE METHOD	465
1.1 Information on Inspection Methods	442	5.1 Magnetizing Field	466
1.2 Electronic References	443	5.2 Continuous versus Noncontinuous Fields	467
1.3 Future NDE Capabilities	443	5.3 Inspection Process	468
2 LIQUID PENETRANTS	445	5.4 Demagnetizing the Part	468
2.1 Penetrant Process	445	6 THERMAL METHODS	468
2.2 Reference Standards	446	6.1 Infrared Cameras	468
2.3 Limitations of Penetrant Inspections	447	6.2 Thermal Paints	469
3 RADIOGRAPHY	448	6.3 Thermal Testing	469
3.1 Generation and Absorption of X-Radiation	448	7 EDDY CURRENT METHODS	469
3.2 Neutron Radiography	450	7.1 Eddy Current Inspection	469
3.3 Attenuation of X-Radiation	450	7.2 Probes and Sensors	474
3.4 Film-Based Radiography	452	8 CASE STUDY OF ADHESIVE BOND NDE	474
3.5 Penetrameter	453	8.1 Introduction	474
3.6 Real-Time Radiography	454	APPENDIX: ULTRASONIC PROPERTIES OF COMMON MATERIALS	476
3.7 Computed Tomography	455	REFERENCES	493
4 ULTRASONIC METHODS	456		
4.1 Sound Waves	457		
4.2 Reflection and Transmission of Sound	458		
4.3 Refraction of Sound	459		
4.4 Inspection Process	461		
4.5 Bond Testers	464		

1 INTRODUCTION

This chapter deals with the nondestructive inspection of materials, components, and structures. The term nondestructive inspection (NDI) or nondestructive evaluation (NDE) is defined as that class of physical and chemical tests that permit the detection and/or measurement of significant properties or the detection of defects in a material without impairing its usefulness. The inspection process is often complicated by the fact that many materials are anisotropic and most NDI

techniques were developed for isotropic materials such as metals. The added complication due to the anisotropy usually means that an inspection is more complicated than it would be with isotropic materials.

Inspection of complex materials and structures is frequently carried out by comparing the expected inspection data with a standard and noting any significant deviations. This means a well-defined standard must be available for calibration of the inspection instrumentation. Furthermore, standards also must contain implanted flaws that mimic those that naturally occur in the material or structure to be inspected. Without a well-defined standard to calibrate the inspection process, the analysis of NDI results can be significantly in error. For example, to estimate the amount of porosity in a cast component from ultrasonic measurements, standard calibration specimens with calibrated levels of porosity must be available to calibrate the instrumentation. Without such standards, estimation of porosity from ultrasonic data is a highly speculative process.

This chapter covers some important and some less well-known NDI tests. Since information on the less frequently used tests is not generally in standard texts, additional sources of information are listed in References at the end of the chapter.

Inspection instrumentation must possess four qualities in order to receive widespread acceptance in the NDI community:

1. *Accuracy.* The instrument must accurately measure a property of the material or structure that can be used to infer either its properties or the presence of flaws.
2. *Reliability.* The instrument must be highly reliable, i.e., it must consistently detect and quantify flaws or a property with a high degree of reliability. If an instrument is not reliable, then it may not detect flaws that can lead to failure of the component or it may indicate the presence of a flaw where none exists. The detection of a phantom flaw can mean that an adequate component is rejected, which is a costly error.
3. *Simplicity.* The most frequently used instruments are those used by factory or repair technicians. The inspection community rarely uses highly skilled operators due to the cost constraints.
4. *Low Cost.* An instrument need not be low cost in an absolute sense. Instead, it must be inexpensive relative either to the value of the component under test or to the cost of a failure or aborted mission. For example, in the aircraft industry as much as 12% of the value of the component may be spent on inspection of a flight-critical aircraft component.

1.1 Information on Inspection Methods

To the engineer confronted by a new inspection requirement, there may arise the question of where to find pertinent information regarding an inspection procedure and its interpretation. Fortunately, many potential sources of information about instrumentation and techniques are available for NDI, and a brief examination of this literature is presented here. Many of these references were generated because of the demands of materials used in flight-critical aerospace structures. In this chapter, we will refer only to scientific and engineering books and journals that one would reasonably expect to find in a well-provisioned library. With the rise of the Internet, there are now many electronic sources of information available on the World Wide Web. These include library catalogues, societal home pages, online journals devoted to inspection, home pages of instrument manufacturers with online demonstrations of their capabilities and inspection services, inspection software, and online forums devoted to solving inspection problems. The References provide many such sources. However, with new electronic sources appearing daily, it is only a brief snapshot of those available at the beginning of the twenty-first

century. For those new to the technology, American Society for Testing and Materials (ASTM) standards are particularly valuable because they give very detailed directions on many NDE techniques. More importantly, they are widely accepted standards for inspections. The References also provide sources for those situations where standard inspection methods are not sufficient to detect the material condition of interest.

General NDE Reference Books

General overviews to NDE techniques are provided in Refs. 1–22. The reader will note that some of these citations are not recent, but they are included because of their value to the engineer who does not possess formal training in the latest inspection technologies. Additionally, some older works were included because of their clarity of presentation, completeness, or usefulness to the inspection of complex structures.

NDE Journals

The periodical literature is often a source of the latest research results for new or modified inspection methodologies.^{5,23–28} Some excellent journals are no longer available but are still a valuable source of information or may contain data available nowhere else. Whenever impossible, World Wide Web addresses are provided to give the reader ready access to this material.

1.2 Electronic References

There are many useful electronic references for those working in NDE technology. Only a few of the many useful sites on the World Wide Web are included here. Many sites contain links to other sites that contain information on a special topic of interest to the reader. Because the Web is constantly being updated, the list in the References represents a very brief snapshot of the information available to the NDE community. Some useful sites associated with government agencies were not included due to space limitations. The Web addresses provided are associated with NDE societies,^{5,6,27,30–37} institutes,^{7,8,10,38–40} government agencies,^{25,38,41} and general interest sites.^{26,33,35,42–44} There are also many references for the reader interested in using or modifying existing NDE techniques.^{31,45–47}

1.3 Future NDE Capabilities

At this point, the reader might be tempted to ask if there are new technologies on the horizon that will enable more cost-effective, anticipatory inspection or monitoring of materials and structures. The answer to this is an emphatic yes. There are new developments in solid-state detectors that should significantly affect both inspection capability and cost. For example, optical and X-ray detectors now give the inspector the ability to rapid scan large areas of structures for defects. Many new developments in these areas are the outgrowth of advances in noninvasive medical imaging. By coupling this technology with computer algorithms that search an image, the inspection of large areas can be automated, providing more accurate inspections with much less operator fatigue. Hopefully this technological advance will remove much of the drudgery of detecting the rather small number of flaws in an otherwise large population of satisfactory components.

The area of data fusion is just beginning to be explored in the NDE field. This means that data collected with one technique can be combined with another technique to detect a range of flaws not detected when either is used independently. Data from several techniques can then be coupled at the basic physics level to provide a more complete description of the microstructural details of a material than is now possible.

Table 1 Capabilities of Common NDI Methods

Method	Typical Flaws Detected	Typical Application	Advantages	Disadvantages
Radiography	Voids, porosity, inclusions, and cracks	Castings, forging, Weldments, and structural assemblies	Detects internal flaws; useful on a wide variety of geometric shapes; portable; provides a permanent record	High cost; insensitive to thin laminar flaws, such as tight fatigue cracks and delaminations; potential health hazard
Liquid penetrants technique	Cracks, gouges, porosity, laps, and seams open to a surface	Castings, forging, weldments, and components subject to fatigue or stress corrosion cracking	Inexpensive; easy to apply; portable; easily interpreted	Flaw must be open to an accessible surface, level of detectability operator dependent
Eddy current inspection	Cracks and variations in alloy composition or heat treatment, wall thickness, dimensions	Tubing, local regions of sheet metal, alloy sorting, and coating thickness measurement	Moderate cost; readily automated; portable	Detects flaws which change conductivity of metals; shallow penetration; geometry sensitive
Magnetic particles method	Cracks, laps, voids, porosity, and inclusions	Castings, forging, and extrusions	Simple; inexpensive; detects shallow subsurface flaws as well as surface flaws	Useful on ferromagnetic materials only; surface preparation required; irrelevant indications often occur; operator dependent
Thermal testing	Voids or disbonds in both metallic and nonmetallic materials, location of hot or cold spots in thermally active assemblies	Laminated structures, honeycomb, and electronic circuit boards	Produces a thermal image that is easily interpreted	Difficult to control surface emissivity and poor discrimination between flaw types
Ultrasonic methods	Cracks, voids, porosity, inclusions and delaminations, and lack of bonding between dissimilar materials	Composites, forgings, castings, and weldments and pipes	Excellent depth penetration; good sensitivity and resolution; can provide permanent record	Requires acoustic coupling to component; slow; interpretation of data is often difficult

Finally, the development of new semiconductor-based devices, microelectromechanical systems (MEMSs), and radio-frequency identification (RFID) allows the implantation of monitoring devices into a material at the time of manufacture to enable real-time structural health monitoring. These devices will permit the inspector to detect and quantify material or structural degradation remotely. This should also enable management of the components and structures for optimum usage over their lifetimes. Remote inspection and tracking of material degradation should reduce the burden of inspection while giving the inspector the ability to examine areas of structure that are now called “hidden.” For more information about this rapidly evolving area the reader is referred to the literature.^{29,33,48–50}

This is a brief review of the commonly used NDI methods listed in Table 1 along with types of flaws that each method detects and the advantages and disadvantages of each technique.

For detailed information regarding the capabilities of any particular method, the reader is referred to the literature. A good place to start any search for the latest NDE technology is the home page of the American Society for Nondestructive Testing.⁶

2 LIQUID PENETRANTS

Liquid penetrants are used to detect surface-connected discontinuities, such as cracks, porosity, and laps, in solid, nonporous materials.⁵¹ The method uses a brightly colored visible or fluorescent penetrating liquid which is applied to the surface of a cleaned part. During a specified “dwell time” the liquid enters the discontinuity and is then removed from the surface of the part in a separate step. The penetrant is drawn from the flaw to the surface by a developer to provide an indication of surface-connected defects. This process is depicted schematically in Figs. 1–4. A penetrant indication of a flaw in a turbine blade is shown in Fig. 5.

2.1 Penetrant Process

Both technical societies and military specifications require a classification system for penetrants. Society documents (typically ASTM E165)⁵¹ categorize penetrants into visible and fluorescent, depending on the type of dye used. In each category, there are three types, depending on how the excess penetrant is removed from the part. These are water washable, postemulsifiable, and solvent removable.

The first step in penetrant testing (PT) or inspection is to clean the part. This critical step is one of the most neglected phases of the PT procedure. Since PT only detects flaws that are open to the surface, the flaw and part surface must be free of dirt, grease, oil, water, chemicals, and other foreign materials that might block the penetrant’s entrance into a defect. Typical cleaning procedures use vapor degreasers, ultrasonic cleaners, alkaline cleaners, or solvents.

After the surface is clean, a liquid penetrant is applied to the part by dipping, spraying, or brushing. In this step, the penetrant on the surface is wicked into the flaw. In the case of tight or narrow surface openings, such as fatigue cracks, the penetrant must be allowed to remain on the part for a minimum of 30 min to completely fill the flaw. High-sensitivity fluorescent dye penetrants are used for this type of inspection.

After the dwell time, excess penetrant is removed by one of the processes mentioned previously. For water-based penetrants an emulsifier is sprayed onto the part and again a dwell time is observed. Water is then used to remove the penetrant from the surface of the part. In some cases, the emulsifier is included in the penetrant, so one only needs to wash the part after the penetrant has had time to penetrate the flaw. These penetrants are therefore called “water washable.” Of course, the emulsifier reduces the brightness of any flaw indication because it dilutes the penetrant. Ideally, only the surface penetrant is removed with the penetrant in the flaw left undisturbed.

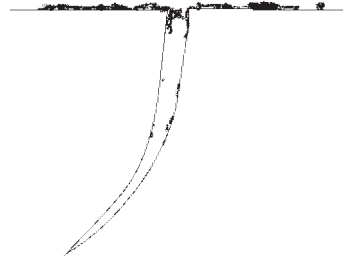


Figure 1 Schematic representation of part surface before cleaning for penetrant inspection.

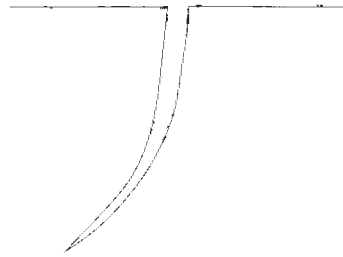


Figure 2 Part surface after cleaning and before penetrant application.

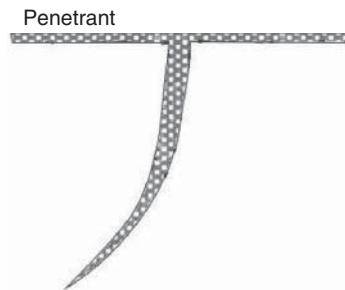


Figure 3 Part after penetrant application.

The final step in a basic penetrant inspection is the application of a fine powder developer. This may be applied either wet or dry. The developer aids in wicking the penetrant from the flaw and provides a suitable background for its detection. The part is then viewed under a suitable illumination—either an ultraviolet or a visible source. A typical fluorescent penetrant indication for a crack in a jet engine turbine blade is shown in Fig. 5.

2.2 Reference Standards

Several reference standards are used to check the effectiveness of liquid penetrant systems. One of the oldest and most often used methods involves applying penetrant to hard chromium-plated brass panels. The panel is bent to place the chromium in tension, producing a series of cracks in

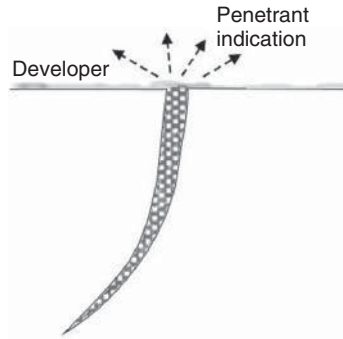


Figure 4 Schematic representation of part after excess penetrant has been removed and developer has been applied.

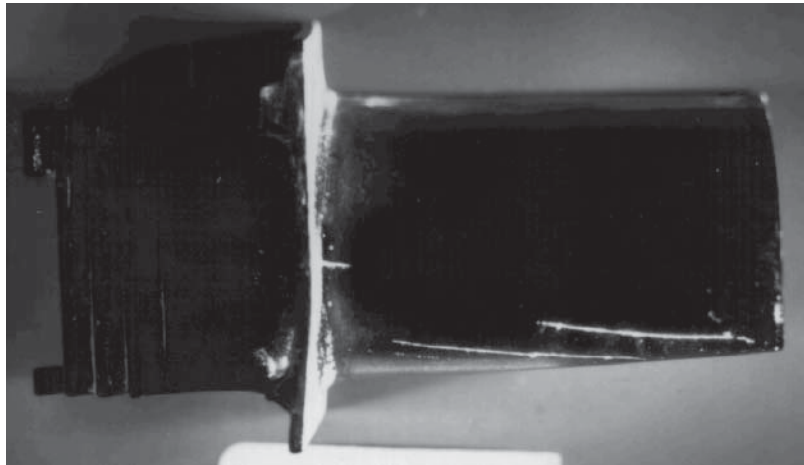


Figure 5 Penetrant indication of crack running along the edge of jet engine turbine blade. Ultraviolet illumination causes the extracted penetrant to fluoresce.

the plating. These panels are available in sets containing fine, medium, and coarse cracks. The panels are used to classify penetrant materials by sensitivity and to detect degrading changes in the penetrant process.

2.3 Limitations of Penetrant Inspections

The major limitation of liquid penetrant inspection is that it can only detect flaws that are open to the surface. Other inspection methods must be used to detect subsurface defects. A factor that may inhibit the effectiveness of liquid penetrant inspection is surface roughness. Rough surfaces are likely to produce false indications by trapping penetrant; therefore, PT is not suited to the inspection of porous materials. Other penetrant-like methods are available for porous components—see the discussion of filtered particle inspection in Ref. 51.

3 RADIOGRAPHY

In radiography used in NDE, the projected X-ray attenuations of a multitude of paths through a specimen are recorded as a two-dimensional image on a recording media, usually film. One might ask if the newer solid-state X-ray imaging technologies used in medicine also apply to NDE. The answer is yes, as will be discussed in the latter part of this section. The most used recording medium is still film because it is the simplest to apply and provides a resolution of subtle details not currently available with solid-state detectors. However, this situation may not be the case much longer as rapid progress is being made in the development of solid-state detectors with significantly enhanced resolution capabilities. Therefore, since this chapter is written at the beginning of the twenty-first century, when film usage for inspection is still commonplace, this portion of the chapter approaches radiography from the standpoint of film-based recording. Since most quantitative relationships for film also apply to solid-state detectors, the material presented should be applicable for the near future.

The radiography testing (RT) process is shown schematically in Figure 6. RT records any feature that changes the attenuation of the X-ray beam as it traverses the component. This local change in attenuation produces a change in the intensity of the X-ray beam, which translates into a change in the density, or darkness, on a film. This change in brightness may appear as a distinct shadow or in some cases a delicate shadow on the radiograph. The inspector is greatly aided in detecting a flaw or discrepancy in a part by his or her knowledge of part shape and its influence on the radiographic image. Flaws, which do not change the attenuation of the X-ray beam on passage through the part, are not recorded. For example, a delamination in a laminated specimen is not visible because there is no local change in attenuation of the X-ray beam as it transverses the part. Conversely, flaws that are oriented parallel to the X-ray path do not attenuate the beam as much, allowing more radiation to expose the film and appearing darker than the surrounding image. An example of a crack in the correct orientation to be visible on a radiograph of a piece of tubing is shown in Fig. 7.

3.1 Generation and Absorption of X-Radiation

X-radiation can be produced from a number of processes. The most common method of generating X rays is with an electron tube in which a beam of energetic electrons impacts a metal target. As the electrons are rapidly decelerated by this collision, a wide band of X-radiation is produced, analogous to white light. This band of radiation is referred to as Bremsstrahlung, or breaking, radiation. These high-energy electrons produce short-wavelength energetic X rays.

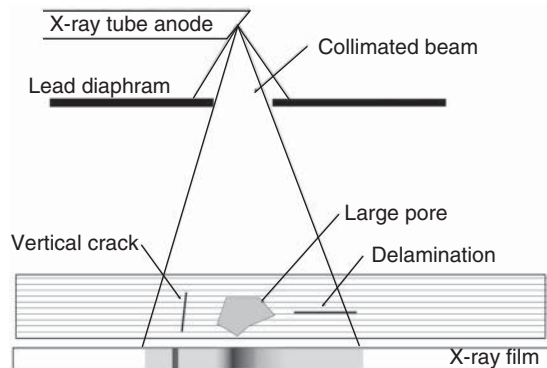


Figure 6 Schematic radiograph with typical flaws.



Figure 7 Radiograph of crack in end of aluminum tubing.

The relationship between the shortest wavelength radiation and the highest voltage applied to the tube is given by

$$\lambda = \frac{12,336}{\text{voltage}}$$

where λ is the wavelength in angstroms and is the shortest wavelength of the X-radiation produced. The more energetic the radiation, the more penetrating powers it possesses, and very high energy radiation is used on dense materials such as metals. While it is possible to analytically predict what X-ray energy would provide the best image for a specific material and geometry, a simpler method of determining the optimum X-ray energy is shown in Fig. 8. Note that high-energy X-ray beams are used for dense materials, e.g., steels, or for thick low-density materials, e.g., large plastic parts. An alternative method to using this figure is to use the radiographic equivalence factors given in Table 2.⁵² Aluminum is the standard material for X-ray tube voltages below 100 KeV, while steel is the standard above this voltage. When radiographing another material, its thickness is multiplied by the factor in this table to obtain the equivalent thickness of the standard material. The radiographic parameters are set up for this thickness of aluminum or steel. When used in this manner, good radiographs can be obtained for most parts. For example, assume that one must radiograph a 0.75-in.-thick piece of brass with a 400-keV X-ray source. The inspector should multiply the 0.75 in. of brass by the factor of 1.3 to obtain 0.98. This means that an acceptable radiograph of the brass plates would be obtained with the same exposure parameters as would be used for 0.98 in. (approximately 1 in.) of steel.

Radiation for RT can also be obtained from the decay of radioactive sources. In this case, the process is usually referred to as gamma radiography. These radiation sources have several

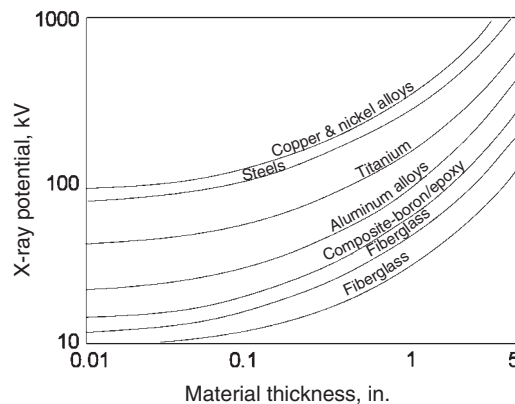


Figure 8 Plot of X-ray tube voltage versus thickness of several industrial materials.

Table 2 Approximate Radiographic Equivalence Factors

Energy Level	100 kV	150 kV	220 kV	250 kV	400 kV	1 MeV	2 MeV	4–25 MeV	¹⁹² Ir	⁶⁰ Co
Magnesium	0.05	0.05	0.08							
Aluminum	0.08	0.12	0.18	—	—	—	—	—	0.35	0.35
Aluminum alloy	0.10	0.14	0.18	—	—	—	—	—	0.35	0.35
Titanium		0.54	0.54	—	0.71	0.9	0.9	0.9	0.9	0.9
Iron/all steels	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Copper	1.5	1.6	1.4	1.4	1.4	1.1	1.1	1.2	1.1	1.1
Zinc		1.4	1.3	—	1.3	—	—	1.2	1.1	1.0
Brass		1.4	1.3	—	1.3	1.2	1.1	1.0	1.1	1.0
Inconel X		1.4	1.3	—	1.3	1.3	1.3	1.3	1.3	1.3
Monel	1.7		1.2							
Zirconium	2.4	2.3	2.0	1.7	1.5	1.0	1.0	1.0	1.2	1.0
Lead	14.0	14.0	12.0	—	—	5.0	2.5	2.7	4.0	2.3
Halfnium			14.0	12.0	9.0	3.0				
Uranium			20.0	16.0	12.0	4.0	—	3.9	12.6	3.4

Source: From Ref. 52.

characteristics that differ from X-ray tubes. First, gamma radiation is very nearly monochromatic; that is, the spectrum of radiation contains only one or two dominant energies. Second, the energies of most sources are on the order of millions-of-volts range, making this source ideal for inspecting highly attenuating materials or very large structures. Third, the small size of these sources permits them to be used in tight locations where an X-ray tube could not fit. Fourth, since the gamma-ray source is continually decaying, adjustments to the exposure time must be made in order to achieve consistent results over time. Finally, the operator must always remember that the source is continually on and is therefore a persistent safety hazard! Aside from these differences, gamma radiography differs little from standard practice, so no further distinction between the two will be given.

3.2 Neutron Radiography

Neutron radiography⁵³ may be useful to inspect some materials and structures. Because the attenuation of neutrons is not related to the elemental composition of the part, some elements can be more easily detected than others. While X rays are most heavily absorbed by high-atomic-number elements, this is not true of neutrons, as shown in Fig. 9. In Fig. 10 two aluminum panels are bonded with an epoxy adhesive. The reader can discern that hydrogen absorbs neutrons more than aluminum does, and thus the missing adhesive is easily detectable.

Neutron radiography, however, does have several constraints. First, neutrons do not expose radiographic film and therefore a fluorescing medium is often used to produce light, which exposes the film. The image produced in this manner is not as sharp and well defined as that from X rays. Second, at present there is no portable high-flux, portable source of neutrons. This means that a nuclear reactor is most often used to supply the neutron radiation. Although neutron radiography has these severe restrictions, at times there is no alternative, and the utility of this method outweighs its expense and complexity.

3.3 Attenuation of X-Radiation

An appreciation of how radiographs are interpreted requires a fundamental understanding of X-ray absorption. The relationship governing this phenomenon is de Beer's law:

$$I = I_0 e^{-\mu x}$$

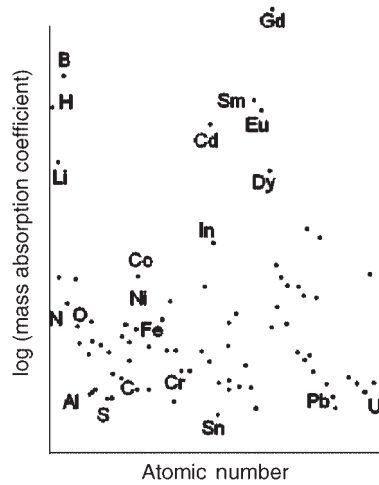


Figure 9 Plot of mass absorption coefficient for neutron radiography versus atomic number.

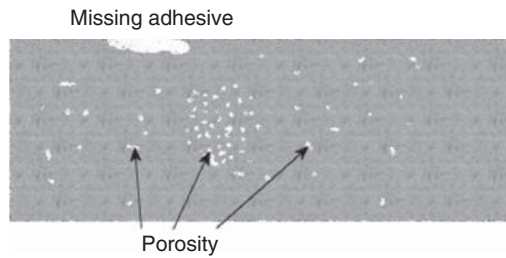


Figure 10 Schematic representation of neutron radiograph showing flaws in adhesive bond.

where I, I_o = transmitted and incident X-ray beam intensities
 μ = attenuation coefficient of material, cm^{-1}
 x = thickness of specimen, cm

Since the attenuation coefficient is a function of both the composition of the specimen and the wavelength or energy of the X rays, it would be necessary to measure it for each wavelength or energy used in RT. However, it is possible to calculate the attenuation coefficient of a material for a specific X-ray energy using the mass absorption coefficients, $(\mu/\rho)_i$, of the elements in the material and their relative abundance in terms of weight percentages. The mass absorption coefficients for most elements are readily available for a variety of X-ray energies⁵⁴. The following example illustrates a method of obtaining the x-ray attenuation coefficient for any compound. Here we have chosen to calculate the attenuation of air to 200 keV x-rays.

$$\mu = (\mu/\rho)_i \rho$$

where μ = is the attenuation coefficient of the compound in cm^{-1}
 $(\mu/\rho)_i$ = is the mass attenuation coefficient in cm^2/gm of each element in the compound
 ρ_i = the density of the compound in gm/cm^3

Using this formula and a simplified composition of the atmosphere, 80% N₂ and 20% O₂, we need only find the mass absorption coefficients of nitrogen and oxygen and then simply multiply these by their relative abundance by weight to find the mass absorption coefficient of the atmosphere.

$$\begin{aligned}(\mu/\rho)_{air} &= 0.80(\mu/\rho)_{nitrogen} + 0.20(\mu/\rho)_{oxygen} \\ &= 0.80 \times 0.598 + 0.20 \times 0.840 \\ &= 0.646 \text{ cm}^2/\text{gm}\end{aligned}$$

Now all that is necessary is to multiply this number by the density of the atmosphere to obtain the attenuation coefficient for air at sea level to 200 keV x-radiation.

$$\begin{aligned}\mu_{air} &= (\mu/\rho)_{air} \times \rho_{air} \\ &= 0.646 \text{ cm}^2/\text{gm} \times 0.001225 \frac{\text{gm}}{\text{cm}^3} \\ &= 7.914 \times 10^{-4}\end{aligned}$$

This procedure is often not used in practice because the results are valid only for a narrow band of wavelengths. Radiographic equivalency factors are used instead. This process points out that each element in a material contributes to the attenuation coefficient by an amount proportional to its amount in the material.

3.4 Film-Based Radiography

The classical method of recording an X-ray image is with film. Because of the continued importance of this medium of recording and the fact that much of the technology associated with it is applicable to newer solid-state recording methods, this section explores film radiography in some detail. The relationship between the darkness produced on an X-ray film and the quantity of radiation impinging on it is shown by log-log plots of darkness, or film density, and relative exposure (Figs. 11 and 12). Varying the time of exposure, intensity of the beam, or specimen thickness changes the density, or darkness, of the image. The slope of the curve along its linear portion is referred to as the film gamma, λ . Film has characteristics that are analogous to

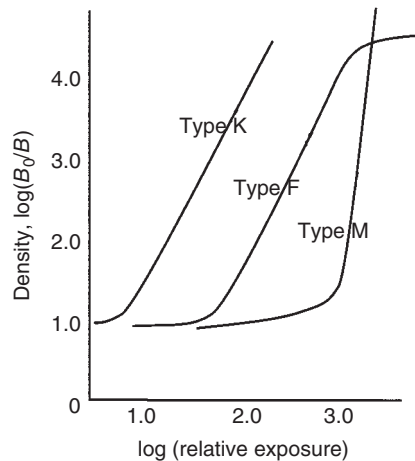


Figure 11 Density or darkness of X-ray film versus relative exposure for three common films.

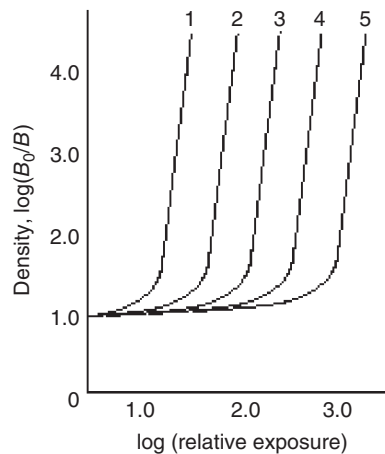


Figure 12 Density versus relative exposure for films that could be used in multiple-film exposure to obtain optimum flaw detectability in a complex part.

electronic devices: The greater the gamma or amplification capability of the film, the smaller its dynamic range—the range of exposures over which density is linearly related to thickness. If it is necessary to use a high-gamma film to detect very subtle flaws in a part with a wide range of thicknesses, then it is necessary to use several different film types in the same cassette or package. In this way, each film will be optimized for flaw detection in a narrow thickness range of the part. Using this information, one may calculate the minimum detectable flaw size for a specific RT inspection. A simple method is available to check the radiographic procedure to determine if this detectability has been achieved on the film. This method does not ensure that the radiograph was taken with the specimen in the proper orientation; it merely provides a method of checking for proper execution of a radiographic procedure; see Section 3.5.

Using knowledge of the minimum density difference that is detectable by the average radiographic inspector, the following equation relates the radiographic sensitivity, S , to radiographic parameters:

$$S = \frac{2.3}{\lambda \mu x}$$

where S is the radiographic sensitivity in percent, λ is the film gamma, μ is the attenuation coefficient of the specimen material, and x is the maximum thickness of the part associated with a particular radiographic film. The radiographer uses a penetrometer to determine if this sensitivity was achieved. Table 3 give the sensitivity S in percent and the expected RT performance in penetrometer values (see Section 3.5).

3.5 Penetrometer

An example of a penetrometer is shown schematically in Fig. 13, while its image on a radiograph is shown in Fig. 14. While there are many types of penetrometers, this one was chosen because it is easily related to radiographic sensitivity. The penetrometer is simply a thin strip of metal or polymeric material⁵⁵ in which three holes of varying sizes are drilled or punched. It is composed of the same material as the specimen and has a thickness 1, 2, or 4% of maximum part thickness. The holes in the penetrometer have diameters that are $1T$, $2T$, and $4T$. The sensitivity achieved for each radiograph is determined by noting the smallest hole just visible in the thinnest penetrometer on a film and using Table 3 to determine the sensitivity achieved.

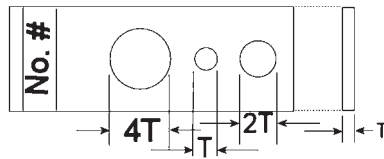


Figure 13 Schematic of typical film penetrometer.

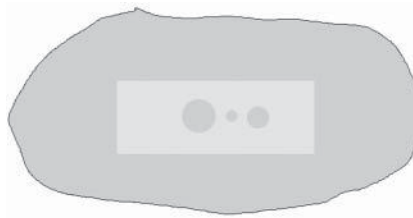


Figure 14 Radiograph of penetrometer shown in Fig. 13. The $1T$ hole is just visible indicating the resolution obtained in the radiograph.

By calculating the radiographic sensitivity and then noting the level achieved in practice, the radiographic process can be quantitatively evaluated. While this procedure does not offer any guarantee of flaw detection, it is useful in evaluating the effectiveness of the RT process.

Almost all variables of the radiographic process may be easily and rapidly changed with the aid of tables, graphs, and nomograms, which are usually provided by film manufacturers free of charge. For more information, the reader is referred to the commercial literature.

3.6 Real-Time Radiography

While film radiography represents the bulk of radiographic NDE performed at this time (the beginning of the twenty-first century), new methods of both recording the data and analyzing it are coming into widespread usage. For example, filmless radiography (FR) and real-time radiography (RTR) use solid-state detectors and digital signal processing (DSP) software instead of film to record and enhance the radiographic image. These methods have many advantages, along with some disadvantages. For example, FR permits viewing a radiographic image while the specimen is being moved. This often permits the detection of flaws that would normally be

Table 3 Radiographic Sensitivity with Thinnest Penetrometer and Smallest Hole Visible on Radiograph

Sensitivity, S (%)	Quality Level (% T – Hole Diameter)
0.7	1 – $1T$
1.0	1 – $2T$
1.4	2 – $1T$
2.0	2 – $2T$
2.8	2 – $4T$
4.0	4 – $2T$

missed in conventional film radiography because of the limited number of views or exposures taken—remember that the X-ray beam must pass along a crack or void for it to be detectable. Additionally, the motion of some flaws enhances their detectability because they present the inspector with a different image as a function of time. Additionally, image enhancement techniques can now be economically and rapidly applied to these images because of the availability of inexpensive, fast computing hardware. The disadvantage of RTR is its lower resolution compared to film. Typical resolution capabilities of RTR or FR systems are in the range of 4 to perhaps 20 line pairs/mm, while film resolution capabilities are in the range of 10–100 line pairs/mm. This means some very fine flaws may not be detectable with FR and the inspector must resort to film. However, in cases where resolution is not the limiting factor, the benefits of software image enhancement can be significant. While the images on film may also be enhanced using the image processing schemes, they cannot be performed in real or near real time, as can be done with an electronic system.

3.7 Computed Tomography

Another advance in industrial radiography has been the incorporation of computed tomography (CT) into the repertoire of the radiographer. Unfortunately, CT has not been exploited to its fullest extent principally due to the high cost of instrumentation. The capability of CT to link NDE measurements with engineering design and analysis gives this inspection a unique ability to provide quantitative estimates of performance not associated with NDE.

The principal advantage of this method is that it produces an image of a thin slice of the specimen under examination. This slice is parallel to the path of the X-ray beam that passes through the specimen, in contrast to the shadowgraph image produced by traditional radiography shown in Fig. 7. Whereas the shadowgraph image can be difficult to interpret, the CT image does not contain information from planes outside the thin slice.

A comparison between CT and traditional film radiography is best made with images from these two modalities. Figure 7 shows a typical radiograph where one can easily see the image

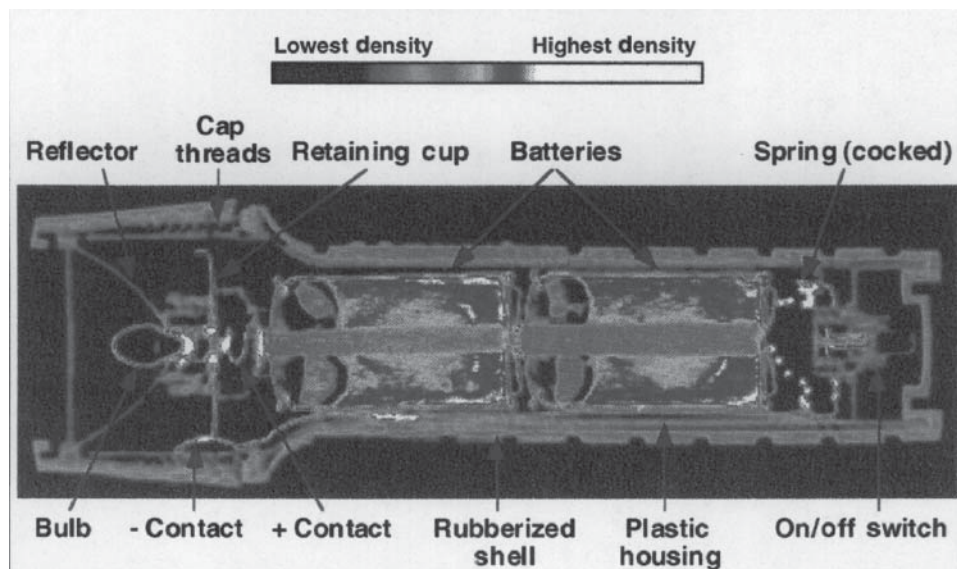


Figure 15 CT image of flashlight showing details of its internal structure.

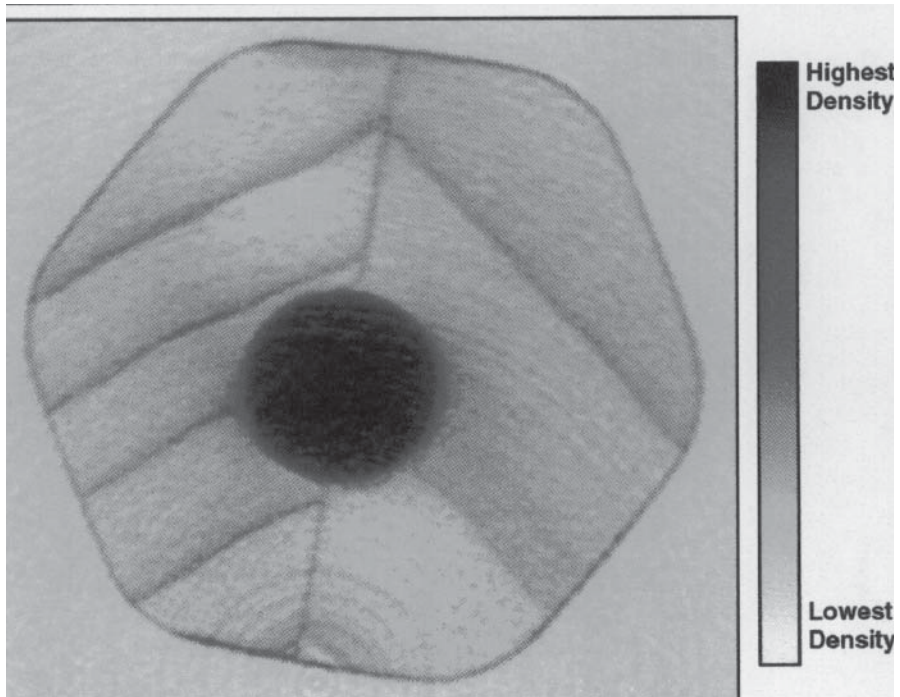


Figure 16 CT image of pencil. The reader will note the yearly growth rings and even the growth variations within a single growing season.

of the top and bottom surfaces of the tube under inspection. The reader can contrast this with the image in Fig. 15, a CT image of a flashlight. The individual components of the flashlight are easily visible and any misplacement of its components or defects in its assembly can be easily detected. An image with a finer scale that reveals the microstructural details of a pencil is shown in Fig. 16. Clearly visible are not only the key features but also the growth rings of the wood. In fact, the details of the growth during each season are visible as rings within rings. The information in the CT image contrasted with conventional radiographs is striking. First, the detectability of a defect is independent of its position in the image. This is not the case with the classical radiograph, where the defect detectability decreases significantly with depth in the specimen, because the defect represents a smaller change in the attenuation of the X-ray beam as the depth increases. Second, the defect detectability is very nearly independent of its orientation. This again is clearly not the case with classical radiography. New applications for CT are constantly being discovered. For example, with a digital CT image it is possible to search for various flaw conditions using computer analysis and relieve the inspector of much of the tedium of examining structures for the odd flaw. In addition, it is possible to link the digital CT image with finite-element analysis software to examine precisely how the flaws present will affect such parameters as stress distribution, heat flow, etc. With little effort, one could analyze the full three-dimensional performance of many engineering structures.

4 ULTRASONIC METHODS

Ultrasonic inspection methods utilize high-frequency sound waves to inspect the interior of solid parts. Sound waves are mechanical or elastic disturbances or waves that propagate in

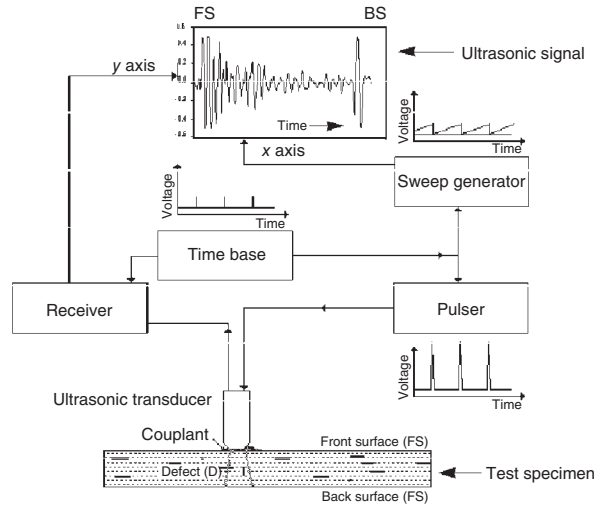


Figure 17 Schematic of ultrasonic data collection and display in A-scan mode.

fluid and solid media. Ultrasonic testing (UT) or inspection is similar to the angler who uses sonar to detect fish.⁵⁶ The government and various technical societies have developed standard practice specifications for UT. These include ASTM specifications 214-68, 428-71, and 494-75 and military specification MIL-1-8950H. Acoustic and ultrasonic testing can take many forms, from simple coin tapping to the transmission and reception of very high frequency or ultrasonic waves into a part to analyze its internal structure. UT instruments operating in the frequency range between 20 and 500 kHz are referred to as sonic instruments, while those that operate above 500 kHz are called ultrasonic. To generate and receive ultrasonic waves, a piezoelectric transducer is employed to convert electrical signals to sound waves and back again. The usual form of a transducer is a piezoelectric crystal mounted in a waterproof housing that is electrically connected to a pulsar (transmitter) and a receiver. In the transmit mode a high-voltage, short-duration electrical spike is applied to the crystal, causing it to rapidly change shape and emit an acoustic pulse. In the receive mode, sound waves or returning echoes compress the piezoelectric crystal, producing an electrical signal that is amplified and processed by the receiver. This process is shown schematically in Fig. 17.

4.1 Sound Waves

Ultrasonic waves have physical characteristics such as wavelength (λ), frequency (f), velocity (v), pressure (p), and amplitude (a). The following relationship between wavelength, frequency, and sound velocity is valid for all sound waves:

$$f\lambda = v$$

For example, the wavelength of longitudinal ultrasonic waves of frequency 2 MHz propagating in steel is 3 mm and the wavelength of shear waves is about half this value, 1.6 mm. The relation between the sound pressure and the particle amplitude is

$$p = 2\pi f a \rho v$$

where f is the frequency of the sound wave, a the amplitude, ρ is density, v is its velocity.

Ultrasonic waves are reflected from boundaries between different materials or media. Each medium has characteristic acoustic impedance and reflections occur in a manner similar to those observed with electrical signals. The acoustic impedance Z (in Rayls or dynes sec/cm³) of any media capable of supporting sound waves is defined by

$$Z = \rho v$$

where ρ = density of medium, g/cm³

v = velocity of sound along direction of propagation

Materials with high acoustic impedance are often referred to as sonically hard, in contrast to sonically soft materials with low impedances. For example, steel ($Z = 7.7 \text{ g/cm}^3 \times 5.9 \text{ km/s} = 45.4 \times 10^6 \text{ kg/m}^2 \cdot \text{s}$) is sonically harder than aluminum ($Z = 2.7 \text{ g/cm}^3 \times 6.3 \text{ km/s} = 17 \times 10^6 \text{ kg/m}^2 \cdot \text{s}$). The appendix at the end of this chapter lists the acoustic properties of many common materials.

4.2 Reflection and Transmission of Sound

Almost all acoustic energy incident on air–solid interfaces is reflected because of the large impedance mismatch between air and most solids. For this reason, a medium with impedance close to that of the part is used to couple the sonic energy from the transducer into the part. A liquid couplant has obvious advantages for parts with a complex geometry, and water is the couplant of choice for most inspection situations. The receiver, in addition to amplifying the returning echoes, also time gates the returning echoes between the front surface and rear surfaces of the component. Thus, any unusually occurring echo is displayed separately or used to set off an alarm, as shown in Fig. 17. This method of displaying the voltage amplitude of the returning pulse versus time or depth (if acoustic velocity is known) at a single point in the specimen is known as an A-scan. In this figure, the first signal corresponds to a reflection from the front surface (FS) of the part and the last signal corresponds to the reflection from its back surface (BS). The signal or echo between the FS and BS is from the defect in the middle of the part.

The portion of sound energy that is either reflected from or transmitted through each interface is a function of the impedances of the medium on each side of that interface. The reflection coefficient R (ratio of the sound pressures or intensities of the reflected and incident waves) and the power reflection coefficient R_{pwr} (ratio of the power in the reflected and incident sound waves) for normally incident waves onto an interface are given as

$$R = \frac{P_r}{P_i} = \frac{Z_1 - Z_2}{Z_1 + Z_2}$$

$$R_{\text{pwr}} = \frac{I_r}{I_i} = \frac{Z_1 - Z_2^2}{Z_1 + Z_2}$$

Likewise, the transmission coefficients T and T_{pwr} are defined as

$$T = \frac{P_t}{P_i} = \frac{2Z_2}{Z_1 + Z_2}$$

$$T_{\text{pwr}} = \frac{I_t}{I_i} = \frac{4(Z_2 \div Z_1)}{[1 + (Z_2 \div Z_1)]^2}$$

where I_i , I_r , and I_t are the incident, reflected, and transmitted acoustic field intensities, respectively; Z_1 is the acoustic impedance of the medium from which the sound wave is incident; and Z_2 is impedance of the medium into which the wave is transmitted. From these equations one can calculate the reflection and transmission coefficients for a planar flaw containing air,

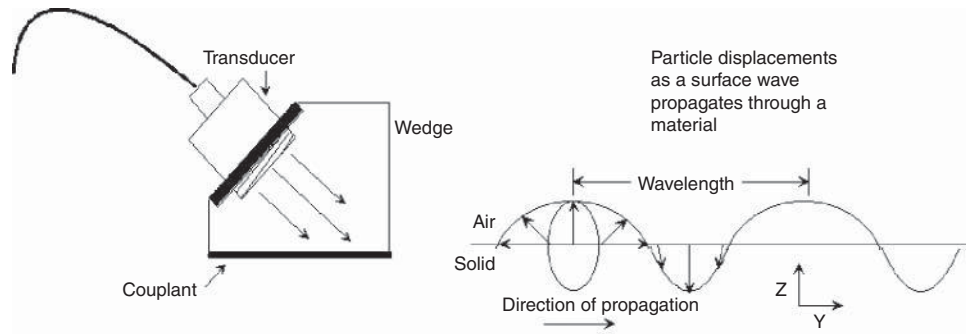


Figure 18 Generation and propagation of surface waves in a material.

$Z_1 = 450 \text{ kg/cm}^2 \cdot \text{s}$, located in a steel part, $Z_2 = 45.4 \times 10^6 \text{ kg/m}^2 \cdot \text{s}$. In this case, the reflection coefficient for the flaw is virtually -1.0 . The minus sign indicates a phase change of 180° for the reflected pulse (note that the defect signal in Fig. 17 is inverted or phase shifted by 180° from the FS signal). Effectively no acoustic energy is transmitted across an air gap, necessitating the use of water as a coupling media in ultrasonic testing. Using the acoustic properties of common materials given in the Appendix the reader can make a number of simple, yet informative, calculations.

Thus far, our discussion has involved only longitudinal waves. This is the only wave that travels through fluids such as air and water. The particle motion in this wave, if one could see it, is similar to the motion of a spring, or a Slinky toy, where the displacement and wave motion are collinear (the oscillations occur along the direction of propagation). The wave is called compressional or dilatational since both compressional and dilatational forces are active in it.

Audible sound waves that we hear are compressional waves. This wave propagates in liquids and gases as well as in solids. However, a solid medium can also support additional types of waves such as shear and Rayleigh or surface waves. Shear or transverse waves have a particle motion that is analogous to what one sees in an oscillating rope. That is, the displacement of the rope is perpendicular to the direction of wave propagation. The velocity of this wave is about half that of compressional waves and is only found in solid media, as indicated in the Appendix. Shear waves are often generated when a longitudinal wave is incident on a fluid–solid interface at angles of incidence other than 90° . Rayleigh or surface waves have elliptical wave motion, as shown in Fig. 18, and penetrate the surface for about one wavelength; therefore, they can be used to detect surface and very near surface flaws. The velocity of Rayleigh waves is about 90% of the shear wave velocity. Their generation requires a special device, or wedge, as shown in Figure 18, which enables an incident ultrasonic wave on the sample at a specific angle that is characteristic of the material (Rayleigh angle). The reader can find more details in the scientific literature.^{57–59}

4.3 Refraction of Sound

The direction of propagation of acoustic waves is governed by the acoustic equivalent of Snell's law. Referring to Fig. 19, the direction of propagation is determined with the equation

$$\frac{\sin \theta_i}{c_I} = \frac{\sin \theta_r}{c_I} = \frac{\sin \gamma_i}{b_I} = \frac{\sin \theta_t}{c_{II}} = \frac{\sin \gamma_t}{b_{II}}$$

where c_I is the velocity of the incident longitudinal wave, c_I and b_I are the velocities of the longitudinal and shear reflected waves. c_{II} and b_{II} are the velocities of the longitudinal and

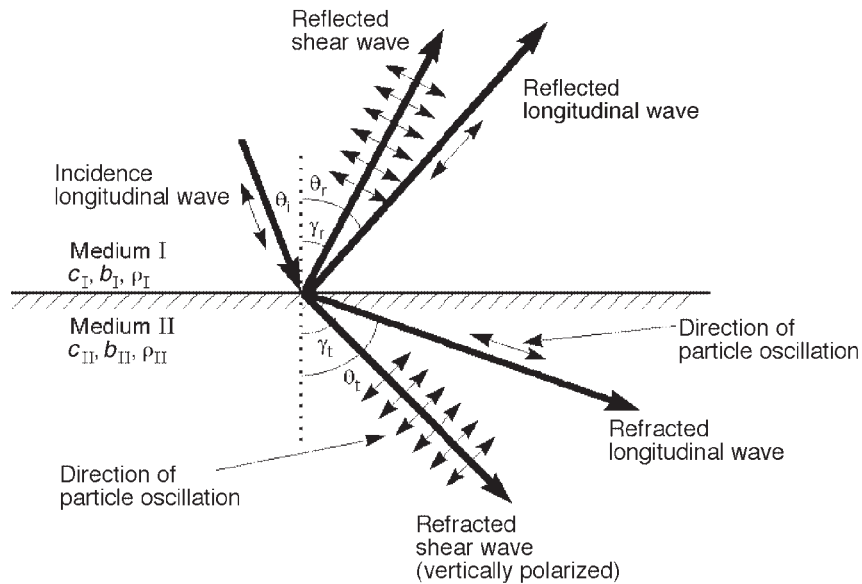


Figure 19 Schematic representation of Snell's law and mode conversion of longitudinal wave incident on solid–solid interface.

shear transmitted waves in solid II. In the water–steel interface, there is no reflected shear wave because these waves do not propagate in fluids such as water. In this case, the above relationship is simplified. Since the water has a lower longitudinal wave speed than either the longitudinal or shear wave speeds of the steel, the transmitted acoustic waves are refracted away from the normal. If the incident wave approaches the interface at increasing angles, there will be an angle above which there will be no transmitted acoustic wave in the higher wave speed material. This angle is referred to as a critical angle. At this angle, the refracted wave travels along the interface and does not enter the solid. A computer-generated curve is shown in Fig. 20 in which the normalized acoustic energy that is reflected and refracted at a water–steel interface is plotted as a function of the angle of the incident longitudinal wave. Note that a longitudinal or first critical angle for steel occurs at 14.5° . Likewise, the shear or second critical angle occurs at about 30° . If the angle of incidence is increased above the first critical angle but less than the second critical angle, only the shear wave is generated in the metal and travels at an angle of refraction described by Snell's law. Angles of incidence above the second critical angle produce a complete reflection of the incident acoustic wave; that is, no acoustic energy enters the solid. At a specific angle of incidence (Rayleigh angle) surface acoustic waves are generating on the material. The Rayleigh angle can be easily calculated from Snell's law by assuming that the refracted angle is 90° . The Rayleigh angle for steel occurs at 29.5° . In the region between the two critical angles, only the shear wave is generated and is referred to as shear wave testing. There are two distinct advantages to inspecting parts with this type of shear wave. First, with only one type of wave present, the ambiguity that would exist concerning which type of wave is reflected from a defect does not occur. Second, the lower wave speed of the shear wave means that it is easier to resolve distances within the part. For these reasons, shear wave inspection is often chosen for inspection of thin metallic structures such as those in aircraft.

Using the relationships for the reflection and transmission coefficients, a great deal of information can be deduced about any ultrasonic inspection situation when the acoustic wave is incident at 90° to the surface. For other angles of incidence, computer software is often used

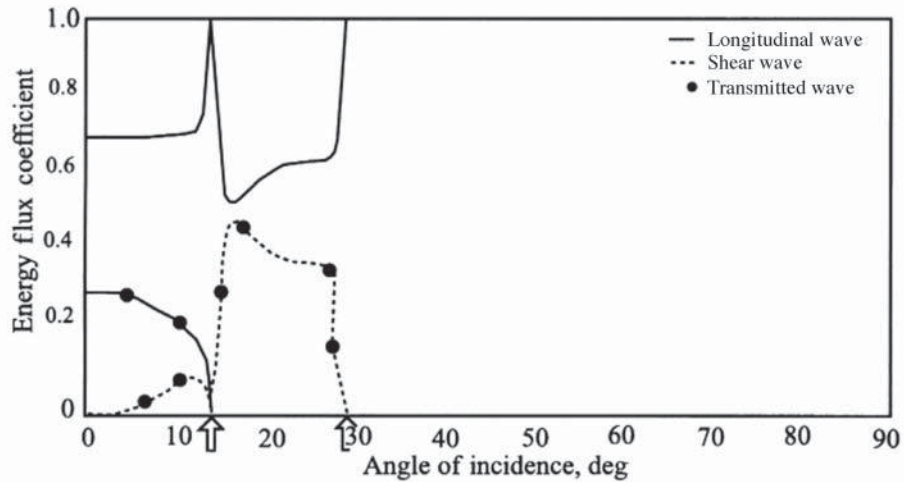


Figure 20 Amplitude (energy flux) and phase of reflected coefficient and transmitted amplitude versus angle of incidence for longitudinal wave incident on water–steel interface. The arrows indicate the critical angles for the interface.

to analyze the acoustic interactions. Analytic predictions of ultrasonic performance in complex materials such as fiber-reinforced composites require the use of more complex algorithms, because more complicated modes of wave propagation can occur. Examples of these include Lamb waves (plate waves), Stoneley waves (interface waves), Love waves (guided in layers of a solid material coated onto another one), and others.

4.4 Inspection Process

Once the type of ultrasonic inspection has been chosen and the optimum experimental parameters determined, one must choose the mode of presentation of the data. If the principal dimension of the flaw is less than the diameter of the transducer, then the A-scan method may be chosen, as shown in Fig. 17. The acquisition of a series of A-scans obtained by scanning the transducer in one direction across the specimen and displaying the data as distance versus depth is referred to as a B-scan. This is the mode most often used by medical ultrasound instrumentation. In the A-scan mode, the size of the flaw may be inferred by comparing the amplitude of the defect signal to a set of standard calibration blocks. Each block has a flat bottom hole (FBH) drilled from one end. Calibration blocks have FBH diameters that vary in $\frac{1}{64}$ -in. increments, for example, a number 5 block has a $\frac{5}{64}$ -in. FBH. By comparing the amplitude of the signal from a calibration block with one from a defect, the inspector may specify a defect size as equivalent to a certain size FBH. The equivalent size is meaningful only for smooth flaws that are nearly perpendicular to the path of the ultrasonic beam and is used in many industrial situations where a reference size is required by a UT procedure. If the flaw size is larger than the transducer diameter, then the C-scan mode is usually selected. In this mode, shown in Fig. 21, the transducer is rastered back and forth across the part. In normal operation, a line is traced on a computer monitor or piece of paper. When a flaw signal is detected between the front and back surfaces, the line drawing ceases and a blank place appears on the paper or monitor. Using this mode of presentation, a planar projection of each flaw is presented to the viewer and its positional relationship to other flaws and to the component boundaries is easily ascertained.

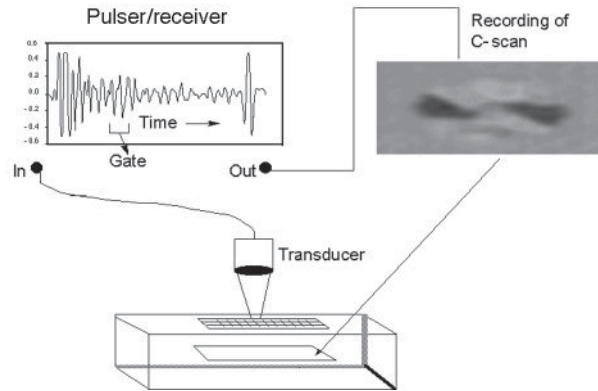


Figure 21 Schematic representation of ultrasonic data collection. The data are displayed using the C-scan mode. The image shows a defect located at a certain depth in the material.

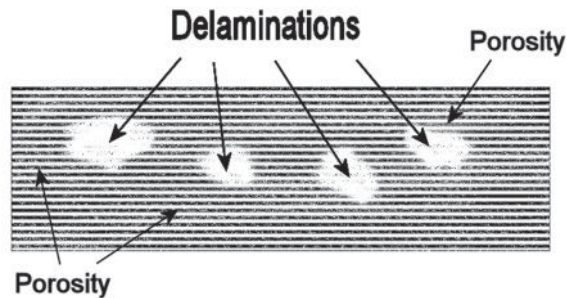


Figure 22 Typical C-scan image of composite specimen showing delaminations and porosity.

Unfortunately, the C-scan mode does not show depth information, unless an electronic gate is set to capture only information from within a specified time window or time gate in the part. With current computer capability, it is a rather simple matter to store all of the returning A-scan data and display only the data in a C-scan mode for a specific depth.

Depending on the structural complexity and the attenuation of the signal, cracklike flaws as small as 0.015 in. in diameter may be reliably detected and quantified with this method. An example of a typical C-scan printout of an adhesively bonded test panel is shown in Fig. 22. This panel was fabricated with a void-simulating Teflon implant and the numerous additional white areas indicate the presence of a great deal of porosity in the part.

Through Transmission versus Pulse Echo

Thus far, the discussion of ultrasonic inspection methods has been concerned with the setup that uses a single transducer to send a signal into the part and to receive any returning echoes. This method is variously referred to as pulse-echo or pitch-catch inspection and is shown schematically in Fig. 23. The other frequently used inspection setup for many structures is called through transmission. With this setup two transducers are used, one to send ultrasonic pulses and the other placed on the opposite side of the part to receive the transmitted signals, as shown schematically in Fig. 24. In Figs. 23 and 24 a large number of reflections occur for

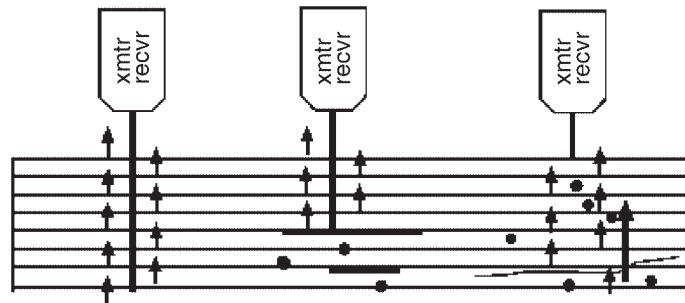


Figure 23 Schematic representation of pulse-echo mode of ultrasonic inspection.

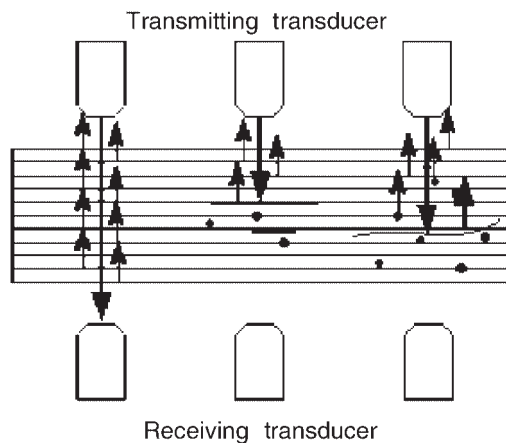


Figure 24 Schematic representation of through-transmission mode of inspection.

the many individual layers in a composite part which can obscure subtle reflections from inclusions whose reflectivity is similar to that of the layered materials. An inclusion with an acoustic impedance very close to that of the part, e.g., paper or peel-plys in polymer-based composites, is very difficult to detect with the through-transmission mode of inspection. In this case, the pulse-echo inspection mode is often used to detect these flaws. On the other hand, reflections from distributed flaws such as porosity, as shown on the right-hand side of Figs. 23 and 24, can be obscured by the general background noise present in an acoustic signal. Therefore, it is the loss in signal strength of the transmitted signal of the through-transmission method that is most often used to detect this type of flaw. While porosity is detectable in this manner, its location may not be determined. In this situation, the pulse-echo mode is required because the distance from the front or back surface to the flaw can be determined by the relative position of the reflections of the scattered porosity with respect to the surface reflection. Because each method supplies important information about potential flaws and its location, modern ultrasonic instrumentation is frequently equipped to perform both types of inspection nearly simultaneously.^{60,61} In such a setup, two transducers are used to conduct a through-transmission test and then each is used separately to conduct pulse-echo tests from opposite sides of the part. This method also helps ensure that a large flaw does not shadow a smaller one, as shown in Figs. 23 and 24.

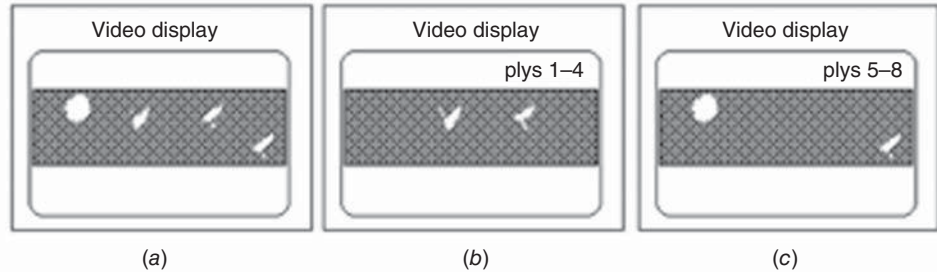


Figure 26 Three different displays of delaminations in 16-ply composite obtained from field-level C-scan system: (a) projection of all flaws through specimen; (b, c) images from selected depths within specimen.

to concerns about the viability of bonded patches on our aging aircraft.⁶² For an extensive treatment of most of the currently used instruments, the reader is referred to review articles.^{63–67} However, while there may seem to be a large number of instruments, some with exaggerated claims of performance, most operate on the same physical principles.

Bond-testing instruments use a variety of means to excite sonic or low-frequency sound waves into the part. In these methods, a low-frequency acoustic transducer is attached to the structure through a couplant. As the driving frequency of the transducer is varied, the amplitude and phase of the transducer oscillations change dramatically as it passes through a resonance. The phase and amplitude of these vibrations change very rapidly and reach a maximum as the driving frequency passes through the resonance frequency of the transducer. The effect of the structure is to dampen the resonant response of the transducer–block combination because of the transfer of acoustic energy into it. Defects such as delaminations and porosity in the adhesive bond layer increase the stiffness of the structure and lower the resonant frequency of the combination. The amplitude of the resonance is increased since there is less material to adsorb the sound energy. These changes in the sharpness of the resonant response are easily detectable electronically.

An alternate method of detecting flaws in bonded components is with a low-frequency or sonic instrument that senses the change in the time of flight for sound waves in the layered structure due to the presence of planar delamination. In such instruments, the increased time of traveling from a transmitting to a receiving transducer is detected electronically. Several commercially available bond-testing instruments successfully exploit this principle. A clever adaptation of a commercial version of this instrument has recently been used to successfully test the joints of structures made from sheet molding compounds.⁶⁴

Probably the most often used method of detecting delaminations in laminated structures is with a coin or tap hammer. This simple instrument is surprisingly effective in trained hands at detecting flaws since an exceedingly complex computer interprets the output signal, i.e., the human brain. Consider, for a moment, that most parents can easily hear their child playing a musical instrument at a school concert. They can perform this task even though their child may have a minor part to play and all the other instruments are much louder than the one that their child is playing. With this powerful real-time signal-processing capability, inspectors can often detect flaws that cannot be detectable with current instrumentation and computers.

5 MAGNETIC PARTICLE METHOD

The magnetic particle method of nondestructive testing is used to locate surface and subsurface discontinuities in ferromagnetic materials.⁶⁸ An excellent reference for this NDE method



Figure 27 Schematic representation of magnetic lines of flux in ferromagnetic metal near a flaw. Small magnetic particles are attracted to the leakage field associated with the flaw.

is Ref. 21, especially Chapters 10–16. Magnetic particle inspection is based on the principle that magnetic lines of force, when present in a ferromagnetic material, are distorted by changes in material continuity, such as cracks or inclusions, as shown schematically in Fig. 27. If the flaw is open to the surface or close to it, the flux lines escape the surface at the site of the discontinuity. Near-surface flaws, such as nonmagnetic inclusions, cause the same bulging of the lines of force above the surface. These distorted fields, usually referred to as leakage fields, reveal the presence of the discontinuity when fine magnetic particles are attracted to them during magnetic particle inspection. If these particles are fluorescent, their presence at a flaw will be visible under ultraviolet light, much like penetrant indications. Magnetic particle inspection is used for steel components because it is fast and easily implemented and has rather simple flaw indications. The part is usually magnetized with an electric current and then a solution containing fluorescent particles is applied by flowing it over the part. The particles stick to the part, forming the indication of the flaw.

5.1 Magnetizing Field

The magnetizing field may be applied to a component by a number of methods. Its function is to generate a residual magnetic field at the surface of the part. The application of a magnetizing force (H) generates a magnetic flux (B) in the component, as shown schematically in Fig. 28. In this figure, the magnetic flux density B has units of Newtons per ampere or Webers per square meter and the strength of the magnetic field or magnetic flux intensity H has units of

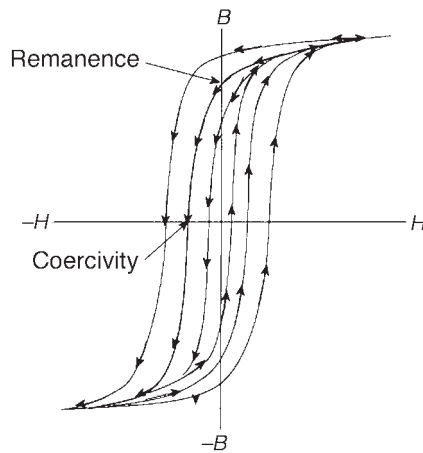


Figure 28 Magnetic flux intensity, H , versus magnetic flux density, B , hysteresis curve for typical steel. Initial magnetization starts at the origin and progresses as shown by the arrows. Demagnetization follows the arrows of the smaller hysteresis loops.

Oersteds or amperes per meter. Starting at the origin, a magnetizing force is applied and the magnetic field internal to the part increases in a nonlinear fashion along the path shown by the arrows. If the force is reversed, the magnetic field does not return to zero but follows the arrows around the curve as shown. The reader will note that once the magnetizing force is removed, the flux density does not return to zero but remains at an elevated value called the material's Remanence. This is the point at which most magnetic particle inspections are performed. The reader will also note that an appreciable reverse magnetic force H must be applied before the internal field density is again zero. This point is referred to as the Coercivity of the material. If the magnetizing force is applied and reversed, the material will respond by continually moving around this hysteresis loop.

Selection of the type of magnetizing current depends primarily on whether the defects are open to the surface or are wholly below it. Alternating-current (AC) magnetization is best for the detection of surface discontinuities because the current is concentrated in the near-surface region of the part. Direct-current (DC) magnetization is best suited for subsurface discontinuities because of its deeper penetration of the part. While DC can be obtained from batteries or DC generators, it is usually produced by half-wave or full-wave rectification of commercial power. Rectified current is classified as half-wave direct current (HWDC) or full-wave direct current (FWDC). Alternating-current fields are usually obtained from conventional power mains, but it is supplied to the part at reduced voltage for reasons of safety and the high-current requirements of the magnetizing process.

Two general types of magnetic particles are available to highlight flaws. One type is low-carbon steel with high-permeability and low-Retentivity particles, which are used dry and consist of different sizes and shapes to respond to leakage fields. The other type is very fine particles of magnetic iron oxide that are suspended in a liquid (either a petroleum distillate or water). These particles are smaller and have a lower permeability than the dry particles. Their small mass permits them to be held by the weak leakage fields at very fine surface cracks. Magnetic particles are available in several colors to increase their contrast against different surfaces or backgrounds. Dry powders are typically gray, red, yellow, and black, while wet particles are usually red, black, or fluorescent.

5.2 Continuous versus Noncontinuous Fields

Because the field is always stronger while the magnetizing current is on, the continuous magnetizing method is generally preferred. Additionally, for specimens with low Retentivity this continuous method is often preferred. In the continuous method, the current can be applied in short pulses, typically 0.5 s. The magnetic particles are applied to the surface during this interval and are free to move to the site of the leakage fields. Liquid suspended fluorescent particles produce the most sensitive indications. For field inspections, the magnetizing current is often continuously applied during the test to give time for the powder to migrate to the defect site. In the residual method, the particles are applied after the magnetizing current is removed. This method is particularly well suited for production inspection of multiple parts.

The choice of direction of the magnetizing field within the part involves the nature of the flaw and its direction with respect to the surface and the major axis of the part. In circular magnetization, the field runs circumferentially around the part. It is induced into the part by passing current through it between two contacting electrodes. Since flaws perpendicular to the magnetizing lines are readily detectable, circular magnetization is used to detect flaws that are parallel or less than 45° to the surface of the long, circular specimens. Placing the specimen inside a coil to create a field running lengthwise through the part produces longitudinal magnetization. This induction method is used to detect transverse discontinuities to the axis of the part.

5.3 Inspection Process

The surface of the part to be examined should be clean, dry, and free of contaminants such as oil, grease, loose rust, loose sand, loose scale, lint, thick paint, welding flux, and weld splatter. Cleaning of the specimen may be accomplished with detergents, organic solvents, or mechanical means, such as scrubbing or grit blasting.

Portable and stationary equipment is available for this inspection process. Selection of the specific type of equipment depends on the nature and location of testing. Portable equipment is available in lightweight units (35–90 lb) which can be readily taken to the inspection site. Generally, these units operate at 115, 230, or 460 V AC and supply current outputs of 750–1500 A in HWAC.

5.4 Demagnetizing the Part

Once the inspection process is complete, the part must be demagnetized. This is done by one of several ways depending on the subsequent usage of the component. A simple method of demagnetizing to remove residual magnetism from small tools is to draw it through the loop-shaped coil tip of a soldering iron. This has the effect of retracing the hysteresis loop a large number of times, each time with a smaller magnetizing force. When completely withdrawn, the tool will then have a very small remnant magnetic field, which for all practical purposes is zero. This same process is accomplished with an industrial part by slowly reducing and reversing the magnetizing current until it is essentially zero, as shown schematically by the arrows in Fig. 28. Another method of demagnetizing a part is to heat it above its Curie temperature (about 550°C for iron), at which point all residual magnetism disappears. This last process is the best means of removing all residual magnetism, but it requires the expense and time of an elevated heat treatment.

6 THERMAL METHODS

Thermal nondestructive inspection methods involve the detection of infrared energy emitted from the surface of a test object.⁶⁹ This technique is used to detect the flow of thermal energy either into or out of a specimen and the effect of anomalies on the surface temperature distribution. The material properties that influence this method are heat capacity, density, thermal conductivity, and emissivity. Defects that are usually detected include porosity, cracks, and delaminations that are parallel to the surface. The sensitivity of any thermal method is greatest for near-surface flaws that impede heat flow and degrades rapidly for deeply buried flaws in high-conductivity materials. Materials with lower thermal conductivity yield better resolution because they allow larger thermal gradients.

6.1 Infrared Cameras

All objects emit infrared (IR) radiation with a temperature above absolute zero. At room temperature, the thermal radiation is predominately IR with a wavelength of approximately 10 μm . IR cameras are available that can produce images from this radiation and are capable of viewing large areas by scanning. Since the IR images are usually captured and stored in digital form, image processing is easily performed and the enhanced images are stored on magnetic or optical media. For many applications, a non-calibrated thermal image of a specimen is sufficient to detect flaws. However, if absolute temperatures are required, the IR instrumentation must be calibrated to account for the surface emissivity of the test subject.

The ability of thermography to detect flaws is often affected by the type of flaw and its orientation with respect to the surface of the object. To have a maximum effect on the surface

temperature, the flaw must interrupt heat flow to the surface. Since a flaw can occur at any angle to the surface, the important parameter is its projected area to the camera. Subsurface flaws such as cracks parallel to the surface of the object, porosity, and debonding of a surface layer are easily detected. Cracks that are perpendicular to the surface can be very difficult or impossible to detect using thermography.

Most thermal NDE methods do not have good spatial resolution due to spreading of thermal energy as it diffuses to the surface. The greatest advantage of thermography is that it can be a noncontact, remote-viewing technique requiring only line-of-sight access to one side of a test specimen. Large areas can be viewed rapidly, since scan rates for IR cameras run between 16 and 30 frames per second. Temperature differences of 0.02°C or less can be detected in a controlled environment.

6.2 Thermal Paints

A number of contact thermal methods are available for inspection purposes. These usually involve applying a coating to the sample and observing a color change as the specimen is thermally cycled. Several different types of coatings are available that cover a wide temperature ranges. Temperature-sensitive pigments in the form of paints have been made to cover a temperature range of 40–1600°C. Thermal phosphors emit visible light when exposed to UV radiation. (The amount of visible light is inversely proportional to temperature.) Thermochromic compounds and cholesteric liquid crystals change color over large temperature ranges. The advantages of these approaches are the simplicity of application and relatively low cost if only small areas are scanned.

6.3 Thermal Testing

Excellent results may be achieved for thermographic inspections performed in dynamic environments where the transient effects of heat flow in the test object can be monitored. This enhances detection of areas where different heat transfer rates are present. Applications involving steady-state conditions are more limited. Thermography has been successfully used in several different areas of testing. In medicine, it is used to detect subsurface tumors. In aircraft manufacturing and maintenance, it may be used to detect debonding in layered materials and structures. In the electronics industry, it is used to detect poor thermal performance of circuit board components. Recently thermography has been used to detect stress-induced thermal gradients around defects in dynamically loaded test samples. For more information on thermal NDE methods, the reader is referred to Refs. 69-71.

7 EDDY CURRENT METHODS

7.1 Eddy Current Inspection

Eddy current (EC) methods are used to inspect electrically conducting components for flaws. Flaws that cause a change in electrical conductivity or magnetic permeability of a part such as surface-breaking EC methods are detected using EC methods. Thickness measurements and the thickness of non-conducting coatings on metal substrates can also be determined with EC methods.⁷² Quite often, several of these conditions can be monitored simultaneously if instrumentation capable of measuring the phase of the EC signal is used.

This inspection method is based on the principle that eddy currents are induced in a conducting material when a coil (probe) is excited with an alternating or transient electric current that is placed in close proximity to the surface of a conductor. The induced currents create an electromagnetic field that opposes the field of the inducing coil in accordance with Lenz's law.

The eddy currents circulate in the part in closed, continuous paths, and their magnitude depends on many variables. These include the magnitude and frequency of the current in the inducing coil, the coil's shape and position relative to the surface of the part, electrical conductivity, magnetic permeability, shape of the part, and presence of discontinuities or inhomogeneities within the material. Therefore, EC inspection is useful for measuring the electrical properties of materials and detecting discontinuities or variations in the geometry of components.

Skin Effect

Eddy current inspections are limited to the near-surface region of the conductor by the skin effect. Within the material, the EC density decreases with the depth. The density of the EC field falls off exponentially with depth and diminishes to a value of about 37% of the surface value at a depth referred to as the standard depth of penetration (SDP). The SDP in meters is calculated with the formula

$$\text{SDP} = \frac{1}{\sqrt{\pi f \sigma \mu}}$$

where f = test frequency, Hz

σ = test material electrical conductivity, mho/m

μ = permeability, H/m

The latter quantity is the product of the relative permeability of the specimen, 1.0 for nonmagnetic materials, and the permeability of free space, $4\pi \times 10^{-7}$ H/m.

Impedance Plane

While the SDP is used to give an indication of the depth from which useful information can be obtained, the choice of the independent variables in most test situations is usually made using the impedance plane diagram suggested by Förster.⁷³ It is theoretically possible to calculate the optimum inspection parameters from numerical codes based on Maxwell's equations, but this is a laborious task that is justified in special situations.

The eddy currents induced at the surface of a material are time varying and have amplitude and phase. The complex impedance of the coil used in the inspection of a specimen is a function of a number of variables. The effect of changes in these variables can be conveniently displayed with the impedance diagram, which shows the variations in amplitude and phase of the coil impedance as functions of the dependent variable specimen conductivity, thickness, and distance between the coil and specimen, or lift-off. For the case of an encircling coil on a solid cylinder, shown schematically in Fig. 29, the complex impedance plane is displayed in Fig. 30. The reader will note that the ordinate and abscissa are normalized by the inductive reactance of the empty coil. This eliminates the effect of the geometry of the coil and specimen. The numerical values shown on the large curve, which are called reference numbers, are used to combine the effects of the conductivity, size of the test specimen, and frequency of the measurement into a single parameter. This yields a diagram that is useful for most test conditions. The reference numbers shown on the outermost curve are obtained with the following relationship for nonmagnetic materials:

$$\text{Reference number} = r\sqrt{2\pi f\mu\sigma}$$

where r = radius of bar, m

f = frequency of test, Hz

m = magnetic permeability of free space, $4\pi \times 10^{-7}$ H/m

σ = conductivity of specimen, mho / m

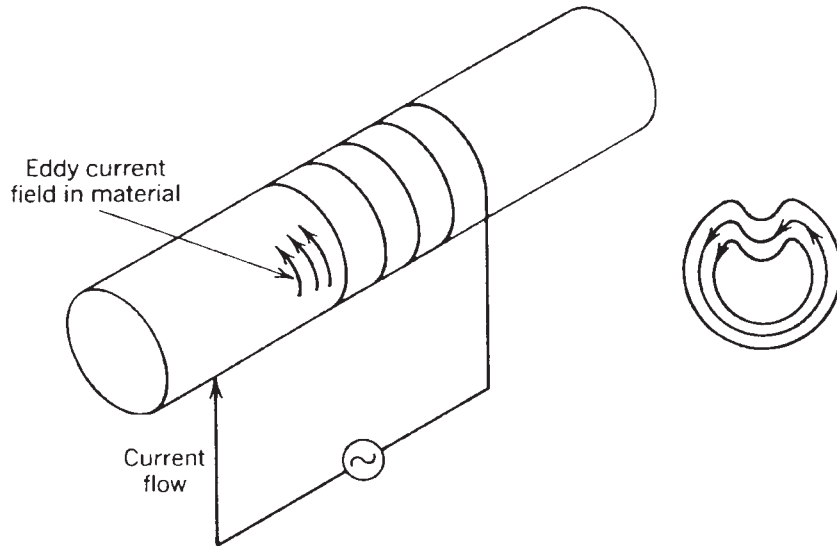


Figure 29 Schematic representations for eddy current inspection of solid cylinder. Also shown are the eddy current paths within the cross section of the cylinder near the crack.

The outer curve in Figs. 30 and 31 is useful only for the case where the coil is the same size as the solid cylinder, which can never happen. For those cases where the coil is larger than the test specimen, which is usually the case, a coil-filling factor is calculated. This is quite easily accomplished with the formula

$$N = \frac{\text{diameter}_{\text{specimen}}}{\text{diameter}_{\text{coil}}}$$

Figure 30 shows the impedance plane with a curve for specimen/coil inspection geometry with a fill factor of 0.75. Note that the reference numbers on the curves representing the different fill factors can be determined by projecting a straight line from the point 1.0 on the ordinate to the reference number of interest, as is shown for the reference number 5.0. Both the fill factor and the reference number change when the size of either the specimen or coil changes. Assume that a reference number of 5.0 is appropriate to a specific test with $N = 1.0$; if the coil diameter is changed so that the fill factor becomes 0.75, then the new reference number will be equal to approximately 7. While the actual change in reference number for this case follows the path indicated by the dashed line in Fig. 30, we have estimated the change along a straight line. This yields a small error in optimizing the test setup but is sufficient for most purposes. For a more detailed treatment of the impedance plane, the reader is referred to Ref. 72. The inspection geometry discussed thus far has been for a solid cylinder. The other geometry of general interest is the thin-walled tube. In this case the skin effect limits the thickness of the metal that may be effectively inspected.

For an infinitely thin-walled tube, the impedance plane is shown in Fig. 31, which includes the curve for a solid cylinder. The dashed lines that connect these two cases are for thin-walled cylinders of varying thicknesses. The semicircular curve for the thin cylinder is used in the same manner as described above for the solid cylinder.

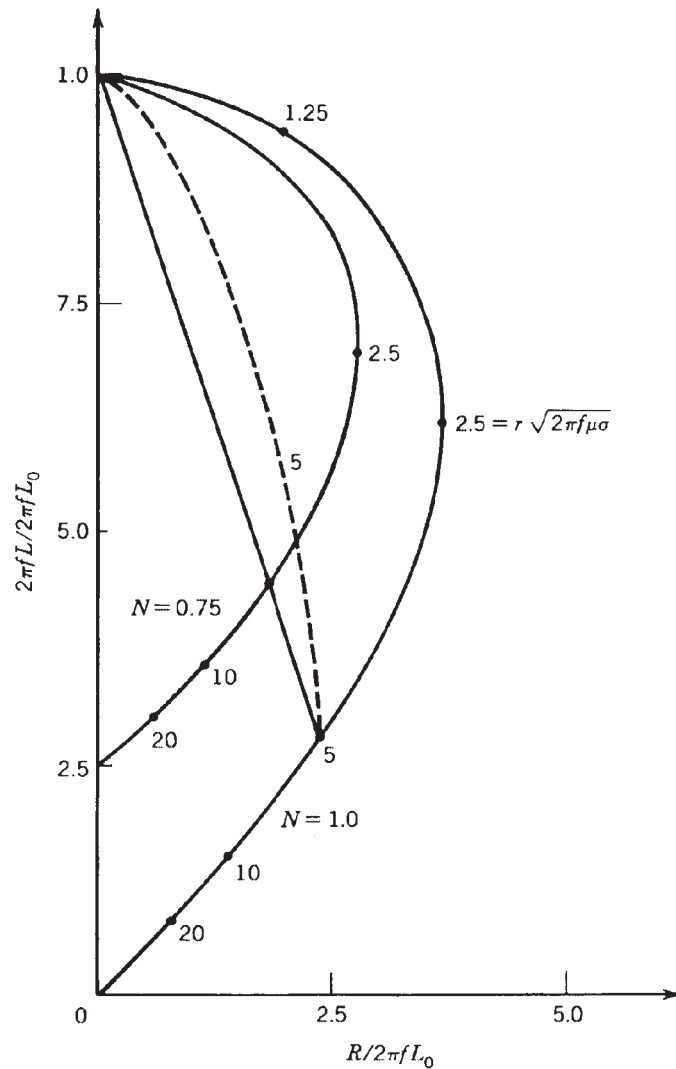


Figure 30 Normalized impedance diagram for long encircling coil on solid, nonferromagnetic cylinder. For $N = 1$ the coil and cylinder have the same diameter, while for $N = 0.75$ the coil is approximately 1.155 times larger than the cylinder.

Lift-Off of Inspection Coil from Specimen

In most inspection situations, the only independent variables are frequency and lift-off. High-frequency excitations are frequently used for detecting defects such as surface-connected cracks or corrosion, while low frequencies are used to detect subsurface flaws. It is also possible to change the coil shape and measurement configuration to enhance detectability, but the discussion of these more complex parameters is beyond the scope of this chapter and the reader is referred to the literature. The relationships discussed so far may be applied by examining Fig. 32, where changes in thickness, lift-off, and conductivity are represented by vectors. These vectors all point in different directions representing the phases of the different

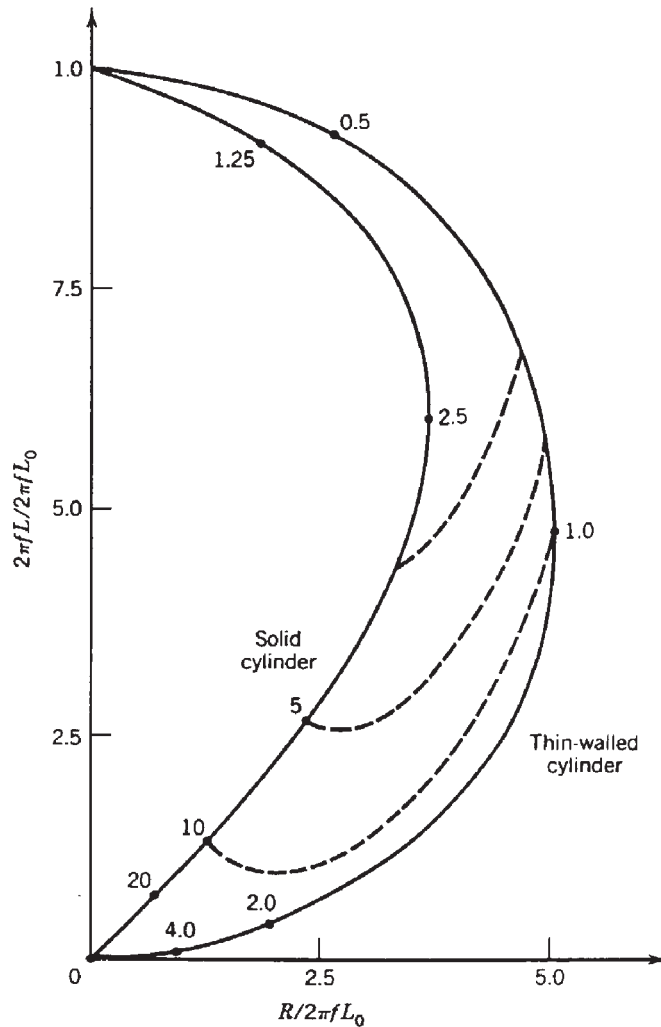


Figure 31 Normalized impedance diagram for long encircling coil on both solid and thin-walled conductive but nonferromagnetic cylinders. The dashed lines represent the effects of varying wall thicknesses.

possible signals. Instrumentation with phase discrimination circuitry can differentiate between these signals and therefore is often capable of detecting two changes in specimen condition at once. Changes in conductivity can arise from several different conditions. For example, aluminum alloys can have different conductivities depending on their heat treatment. Changes in apparent conductivity are also due to the presence of cracks or voids.

A crack decreases the apparent conductivity of the specimen because the eddy currents must travel a longer distance to complete their circuit within the material. Lift-off and wall thinning are also shown in Fig. 32. Thus, two different flaw conditions can be rapidly detected. There are situations where changes in wall thickness and lift-off result in signals that are very nearly out of phase and therefore the net change is not detectable. If this situation is suspected, then inspection at two different frequencies is warranted. There are other inspection situations

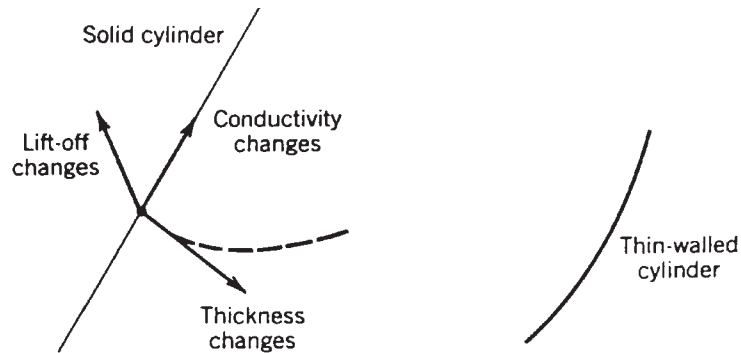


Figure 32 Effects of various changes in inspection conditions on local signal changes in impedance plane of Fig. 31. Phase differentiation is relatively easily accomplished with current instrumentation.

that cannot be covered in this brief description. These include the inspection of ferromagnetic alloys, plate, and sheet stock and the measurement of film thicknesses on metal substrates. For a treatment of these and other special applications of EC inspection, the reader is referred to Ref. 72.

7.2 Probes and Sensors

In some situations, it may be advantageous to have a core with a high magnetic permeability inside the coil. Magnetic fields will pass through the medium with the highest permeability if possible. Therefore, materials with high permeability can be placed in different geometric configurations to enhance the sensitivity of a probe. One example of this is the cup-core probe where the coil has a ferrite core, shield, and cap.⁷⁴ For low-frequency or transient tests, using the inductive coil as a sensor is not sufficient since it responds to the time change in magnetic field and not the direct magnetic field. It is necessary to induce a magnetic field sensor that responds well at the lower frequencies. Sensors such as the Hall effect sensor and giant magneto-resistive (GMR) sensors have been used to accomplish this.⁷⁵ There are numerous methods of making eddy current NDE measurements. Two of the more common methods are shown schematically in Fig. 33. In the absolute coil arrangement, very accurate measurements can be made of the differences between the two samples. In the differential coil method, it is the differences between the two variables at two slightly different locations that may be detected. For this arrangement, slightly varying changes in dimensions and conductivity are not sensed, while singularities such as cracks or voids are highlighted, even in the presence of other slowly changing variables. Since the specific electronic circuitry used to accomplish this task can vary dramatically, depending on the specific inspection situation, the reader is referred to the current NDE and instrumentation in Refs. 73, 76, and 77.

8 CASE STUDY OF ADHESIVE BOND NDE

8.1 Introduction

Usually NDE is one of the last considerations in the design process of a new or modified structure. As such, the NDE specialist must attempt to insert the inspection process into an established design and/or manufacturing process. This has often stymied the development of new approaches to the nondestructive inspection field. This is particularly true for a key joining

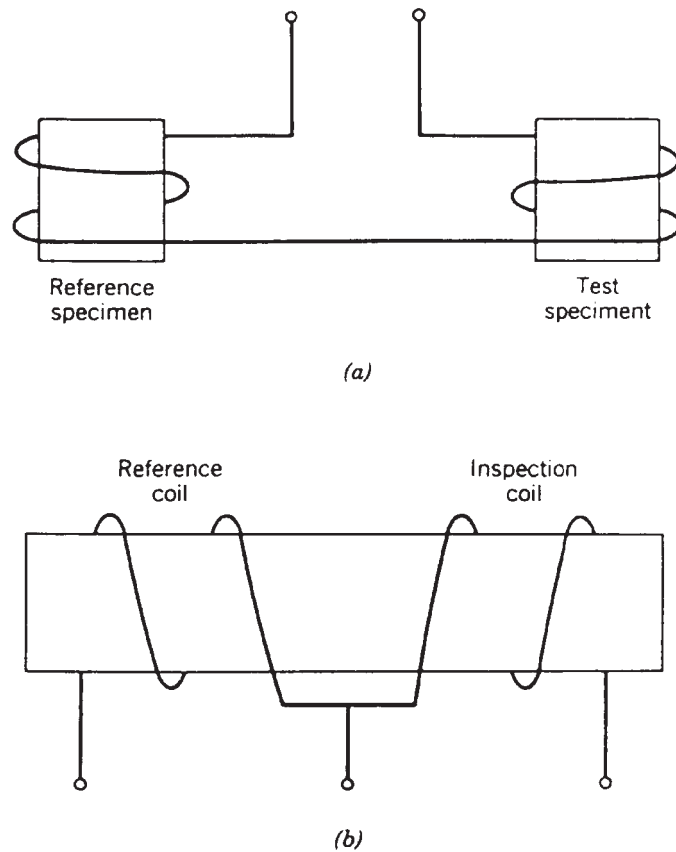


Figure 33 Schematic representation of (a) absolute versus (b) differential coil configurations used in eddy current testing.

technology—adhesive bonding. Because of the limitations imposed by the use of the standard big-5 techniques outlined in this chapter, reliable adhesive bonding was not achieved until very recently. Since the early 1950s the advantages of adhesive bonding in the design and construction of all lightweight structures has been apparent and many research projects were initiated in order to find a method to guarantee a minimum strength of an adhesive bond. The methodology usually chosen was an ultrasonic through transmission inspection of the bond and an attempt to correlate any changes in sound speed or attenuation to the ultimate load carrying capability of the bonded joint. This technique can only be used in the middle of most bonded joints due to diffraction of the sound beam by edges. That meant that the inspection was carried out in an area of the bond that carried no load.⁷⁸ This means that the previous work on ultrasonic correlation to bond strength could at best be fortuitous. During work on laser shock peening during the 1990's it became apparent that it was possible to control laser fluence so that a high strength alloy material could be fractured or subjected to precise tensile loading.^{79,80} This then led to the development of a laser system to provide the designer a precise estimate of the lower bound of bond strength.⁸¹ Unfortunately, this inspection is both expensive and can only be performed after the bonded joint is assembled. For this reason, a preassembly inspection of the bonded surfaces was needed and work began on a technique to provide that capability.

Recent work on a technique to quantify the surfaces of a bonded structure before assembly has been quite successful and is now an accepted technique in several industries where it is desirable to use bonded joints.⁸²⁻⁸⁴

The point of this case study is to demonstrate that adherence to traditional methodologies where the measured parameters (ultrasonic speed and attenuation) are not related to the desired measurement (strength) often leads to erroneous results and little progress. It is necessary to discover a method that measures the desired property that in use does not introduce unacceptable flaws into a material or structure. If a method can be discovered and developed to provide data about the reliability of a material or structure, then a new structural paradigm may be enabled. This is now the case with the widespread use of bonded joints in composite structures not previously possible due to the lack of an NDT capability to ensure the reliability of bonded joints.

APPENDIX: ULTRASONIC PROPERTIES OF COMMON MATERIALS

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Gases							
Alcohol vapor				0.23			
Air	25 atm			0.33			
	50 atm			0.34			
	100 atm			0.35			
				0.34			
				0.39			
				0.55			
Ammonia				0.42			
Argon		0.00178	0.32				
Carbon monoxide		0.34					
Carbon dioxide				0.26			
Carbon disulfide				0.19			
Chlorine				0.21			
Ether vapor				0.18			
Ethylene				0.31			
Helium		0.00018	0.97				
Hydrogen		0.00009	1.28				
Methane		0.00074	0.43				
Neon		0.0009	0.43				
Nitric oxide				0.33			
Nitrogen		0.00125	0.33				
		0.00116	0.35				
Nitrous oxide				0.26			
Oxygen		0.00142	0.32				
		0.00132	0.33				
Water vapor				0.4			
				0.41			
				0.42			

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Gases—Cryogenic							
Argon		1.404	0.84			1.18	
		1.424	0.86			1.23	
Helium		0.125	0.18			0.023	
		0.146	0.23			0.034	
Helium-4	Liquid at 2 K	0.15	0.18			0.027	
Hydrogen	Liquid at 20 K	0.07	1.19			0.08	
		0.355	1.13			0.401	
Nitrogen		0.815	0.87			0.708	
		0.843	0.93			0.783	
Oxygen		1.143	0.97			1.04	
		1.272	1.13			1.44	
		1.149	0.95			1.09	
Liquids							
Acetate butyl (<i>n</i>)		0.871	1.17			1.02	
Acetate ethyl		0.9	1.18			1.06	
Acetate methyl		0.928	1.15			1.07	
Acetate propyl		0.891	1.18			1.05	
Acetone		0.79	1.17			0.92	
		0.791	1.16		0.92		0.0469
Acetonitrile		0.783	1.29			1.01	
Acetonyl acetone		0.729	1.4			1.36	
Acetylene dichloride		1.26	1.02			1.29	
Adiprene	CW-520	0.79	1.68		1.33		0.0469
Alcohol, butyl		0.81	1.24			1	
Alcohol, ethyl		0.789	1.18			0.93	
Alcohol, furfuryl		1.135	1.45			1.65	
Alcohol, isopropyl		0.79	1.17		0.92		0.08
Alcohol, methyl		0.792	1.12			0.89	
Alcohol, propyl (<i>i</i>)		0.786	1.17			0.92	
Alcohol, propyl (<i>n</i>)		0.804	1.22			0.98	
Alcohol, <i>t</i> -amyl		0.81	1.2			0.97	
Alkazene 13		0.86	1.32			1.14	
Analine		1.022	1.69			1.68	
A-Spirit	Ethanol > 96%	0.79	1.18			0.93	
Benzene	C ₆ H ₆	0.87	1.3			1.13	
		0.88	1.31			1.15	
Benzol		0.878	1.33			1.17	
Benzol ethyl		0.868	1.34			1.16	
Bromo-benzene	C ₆ H ₅ Br	1.52	1.17			1.78	
Bromoform		2.89	0.92			2.66	
Butanol	Butyl	0.71	1.27		0.9		
		0.81	1.27			1.03	
Butoxyethanol	(2 <i>n</i> -)		1.31				
<i>tert</i> -Butyl chloride		0.84	0.98			0.82	

Material	Comments	Density (g/cm ³)	V ₁ (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Butylene glycol (2.3)		1.019	1.48			1.51	
Butyrate ethyl		0.877	1.17			1.03	
Carbitol		0.988	1.46			1.44	
Carbon disulfide			1.16				
		1.26	1.15			1.45	
Carbon tetrachloride	CCl ₄	1.595	0.93			1.48	
Cerechlor 42		1.26	1.43			1.8	
Chinolin		1.09	1.57			1.71	
Chlorobenzene	C ₆ H ₅ Cl	1.1	1.3			1.43	
Chloroform		1.49	0.99			1.47	
Chlorohexanol	>98%	0.95	1.42			1.35	
Cyclohexanol		0.962	1.45			1.39	
	Freon		1.2				
	DTE 21 oil		1.39				
	Glycerol		1.52				
Decahydronaphtaline		0.948	1.42			1.39	
	C ₁₀ H ₁₈	0.89	1.42			1.27	
	Paraffin		1.41				
Diacetyl		0.99	1.24			1.22	
Diamine propane	(1.3) >99%	0.89	1.66			1.47	
Dichloro isobutane (1.3)		1.14	1.22			1.39	
Diethylamine	(C ₂ H ₅) ₂ NH	0.7	1.13			0.8	
Diethylene glycol		1.116	1.58			1.76	
Diethyl ketone		0.813	1.31			1.07	
Dimethyl phthalate		1.2	1.46			1.75	
Dioxane		1.033	1.38			1.43	
Diphenyl	Diphenyl oxide		1.5				
Dodecanol		0.83	1.41			1.16	
DTE 21	Mobil		1.39				
DTE 24	Mobil		1.42				
DTE 26	Mobil		1.43				
Dubanol	Shell		1.43				
			1.44				
	Water		1.55				
Ethanol	C ₂ H ₄ OH	0.79	1.13		0.89		0.0421
Ethanol amide		1.018	1.72			1.75	
Ethyl acetate		0.9	1.19			1.07	
Ethylenclycolol	Sodium benzoate		1.67				
Ethylene diamine	>99.5%	0.9	1.69			1.52	
Ethylene glycol	>99.5%	1.11	1.69			1.88	
	1.2-Ethenediol	1.112	1.67			1.86	
		1.113	1.66			1.85	

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
	H ₂ O 1:4		1.6				
	H ₂ O 2:3		1.68				
	H ₂ O 3:2		1.72				
	H ₂ O 4:1		1.72				
Ethyl ether		0.713	0.99			0.7	
Fluorinert	H ₂ O-72	1.68	0.51			0.86	
	FC-104	1.76	0.58			1.01	
	FC-75	1.76	0.59			1.02	
	FC-77	1.78	0.6			1.05	
	FC-43	1.85	0.66			1.21	
	FC-40	1.86	0.64			1.19	
	FC-70	1.94	0.69			1.33	
Fluoro-benzene	C ₆ H ₅ F	1.024	1.18			1.21	
Formamide		1.134	1.62			1.84	
Freon			0.68				
Freon MF 21.1		1.485	0.8			1.19	
Freon TF 21.1		1.574	0.97			1.52	
Furfural		1.157	1.45			1.68	
Gasoline		0.803	1.25			1	
Glycerine	CH ₂ OHCHOHCH ₂ OH	1.23	1.9			2.34	
	Glycerol >98%	1.26	1.88			2.37	
		1.26	1.92			2.42	
Glycerol	Water	1.22	1.88			2.29	
	Butanol		1.45				
	Ethanol		1.52				
			1.56				
	Isopropanol		1.57				
Glycerol trioleate		0.91	1.44			1.31	
Glycol	Polyethylene	1.06	1.62			1.71	
		1.087	1.62			1.75	
	Ethylene	1.108	1.59			1.76	
		1.112	1.67			1.86	
Hexane	(<i>n</i> -)C ₆ H ₁₄	0.659	1.1			0.727	
<i>n</i> -Hexanol		0.819	1.3			1.06	
Honey	Sue Bee Orange	1.42	2.03			2.89	
Iodobenzene	C ₆ H ₅ I	1.183	1.1			2.01	
Isopentane		0.62	0.99			0.62	
Isopropanol			1.14				
Isopropyl alcohol		0.786	1.17			0.92	
	Propylene glycol	0.79	1.14			0.9	0
		0.84	1.21			1.01	0.14
		0.88	1.24			1.09	0.4
		0.88	1.28			1.13	0.23
		0.92	1.3			1.2	0.44
		0.94	1.35			1.27	0.33
		0.96	1.36			1.31	0.53

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
		0.97	1.36			1.32	1.12
		0.99	1.43			1.41	0.47
		1	1.42			1.42	1.08
		1	1.43			1.43	0.67
		1.03	1.48			1.52	1.13
		1.03	1.51			1.56	0.72
		1.04	1.54			1.64	1.2
		1.07	1.58			1.7	0.93
		1.09	1.6			1.75	1.4
		1.13	1.65			1.86	1.68
		1.16	1.75			2.03	2.57
		1.2	1.82			2.19	2.12
		1.25	1.92			2.39	4.55
Jeffox WL-1400			1.53				
Kerosene		0.81	1.32			1.07	
Linalool	Hg	0.884	1.4			1.24	
Mercury	Hg	13.6	1.45			19.7	
Mercury, 20°C			1.42			19.7	
Mesityloxiide		0.85	1.31			1.11	
Methanol	CH ₃ OH	0.796	1.09		0.87	0.87	0.0262
Methyl acetate		0.934	1.21			1.13	
Methylene iodide			0.98				
Methylethyl ketone		0.805	1.21			0.97	
Methyl naphthalene		1.09	1.51			1.65	
Methyl salicylate		1.16	1.38			1.6	
Modinet P40		1.06	1.38			1.47	
Monochlorobenzene		1.107	1.27			1.41	
Morpholine		1	1.44			1.44	
M-xylol		1.107	1.27			1.41	
		1	1.44			1.44	
		0.864	1.32			1.14	
NaK	Mix	0.64	1.66				
		0.713	1.72				
		0.714	1.84				
		0.73	1.95				
		0.736	1.77				
		0.738	1.89				
		0.754	2				
		0.759	1.82				
		0.761	1.99				
		0.778	2.05				
		0.781	1.88				
		0.784	1.99				
		0.801	2.1				
		0.804	1.93				
		0.807	2.04				
		0.825	2.15				
		0.826	1.98				

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
		0.83	2.09				
		0.848	2.2				
		0.849	2.04				
		0.853	2.14				
		0.871	2.25				
		0.876	2.19				
		0.893	2.31				
Nicotine	C ₁₀ H ₁₄ N ₂	1.01	1.49			1.51	
Nitrobenzene		1.2	1.46			1.75	
Nitrogen	N ₂	0.8	0.86			0.68	
Nitromethane		1.13	1.33			1.5	
Oil, baby		0.821	1.43			1.17	
Oil, castor	Jeffox WL-1400		1.52				
	Castor	0.95	1.54			1.45	
	Ricinus oil	0.969	1.48			1.43	
Oil, corn		0.922	1.46			1.34	
Oil, cutting	64 AS (red)		1.4				
Oil, diesel			1.25				
Oil, fluorosilicone	Dow FS-1265		0.76				
Oil, grape seed	Cerechlor	0.92	1.43				
	Castor oil	0.936	1.44			1.35	
Oil, gravity fuel AA		0.99	1.49			1.48	
Oil, linseed		0.922	1.77			1.63	
		0.94	1.46			1.34	
Oil, mineral (heavy)		0.843	1.46			1.37	
Oil, mineral (light)		0.825	1.44			1.19	
Oil, motor (2-cycle)			1.43				
Oil, motor (SAE 20)		0.87	1.74			1.51	
Oil, motor (SAE 30)		0.88	1.7			1.5	
Oil, olive			1.43				
		0.918	1.45			1.32	
		0.948	1.43			1.39	
Oil, paraffin			1.28				
			1.43				
		0.835	1.42			1.86	
Oil, peanut		0.914	1.44			1.31	
		0.936	1.46			1.37	
Oil, safflower		0.92	1.45			1.34	
Oil, silicone	Dow 710 fluid		1.35				
	Silicone 200	0.818	0.96			0.74	
		0.94	0.97			0.91	
		0.972	0.99			0.96	
	30 cP	0.993	0.99			0.983	
		1.1	1.37			1.5	
Oil, soybean		0.93	1.43			1.32	
Oil, sperm		0.88	1.44			1.27	
Oil, sun	Nivea		1.41				

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Oil, sunflower		0.92	1.45			1.34	
Oil, synthetic		0.98	1.27			1.33	
Oil, transformer		0.92	1.39			1.28	
Oil, transmission	Dexron (red)		1.42				
Oil, velocite	Mobil		1.3				
Oil, wheat germ		0.94	1.49			1.39	
Paraffin		1.5	1.5			2.3	
<i>d</i> -Penchone		0.94	1.32			1.24	
Pentane		0.621	1.01			0.63	
	(<i>n</i> -) C ₅ H ₁₂	0.626	1.03			0.64	
Petroleum		0.825	1.29			1.06	
Polypropylene glycol	Polyglycol P-400		1.3				
	Polyglycol P-1200		1.3				
	Polyglycol E-200		1.57				
Polypropylene oxide	Ambiflo		1.37				
Potassium		0.662	1.49				
		0.685	1.55				
		0.707	1.6				
		0.729	1.65				
		0.751	1.71				
		0.773	1.76				
		0.796	1.81				
		0.818	1.86				
Propane diol	(1.3) 97%	1.05	1.62			1.7	
Pyridine		0.982	1.41			1.38	
Sodium		0.759	2.15				
		0.784	2.21				
		0.809	2.26				
		0.833	2.31				
		0.857	2.37				
		0.881	2.42				
		0.904	2.48				
		0.926	2.53				
Solvesso #3		0.877	1.37			0.201	
Sonotrack	Coupling gel	1.4	1.62			1.68	
Span 20			1.48				
Span 85			1.46				
Tallow			0.39				
Tetraethylene glycol		1.12	1.58			1.77	
Tetrahydronaphthaline	(1.2.3.4)	0.97	1.47			1.42	
Trichloroethylene		1.05	1.05			0.41	
Triethylene glycol		1.12	1.61			1.81	
		1.123	1.61			1.98	
Trithylamine	(C ₂ H ₅) ₃ N	0.73	1.12			0.81	
Turpentine		0.87	1.25			1.11	
		0.893	1.28			1.14	
			1.27				

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Ucon 75H450			1.54				
Univis 800		0.87	1.35				1.19
Water-salt solution	10%		1.47				
	15%		1.53				
	20%		1.6				
Water, sea		1.025	1.53			1.57	
		1.026	1.5			1.54	
Water		1	1.51			1.51	
		1	1.55			1.55	
	Propylene glycol	1	1.5			1.5	
		1.01	1.61			1.63	0.021
		1.02	1.69			1.72	0.038
		1.03	1.51			1.56	0.669
		1.03	1.62			1.66	0.213
		1.03	1.69			1.73	0.088
		1.05	1.6			1.69	
		1.06	1.69			1.79	0.059
		1.07	1.58			1.7	1.025
		1.07	1.66			1.78	0.395
		1.07	1.71			1.83	0.174
		1.07	1.73			1.84	0.112
		1.11	1.71			1.89	0.086
		1.11	1.75			1.95	0.321
		1.11	1.76			1.94	0.117
		1.11	1.77			1.97	0.182
		1.12	1.66			1.86	1.744
		1.12	1.71			1.91	0.582
		1.16	1.75			2.03	2.57
		1.16	1.78			2.06	1.242
		1.16	1.8			2.09	0.175
		1.16	1.81			2.09	0.648
		1.16	1.82			2.11	0.241
		1.16	1.82			2.11	0.397
		1.2	1.82			2.19	2.12
		1.2	1.85			2.23	1.469
		1.2	1.85			2.22	2.033
		1.2	1.86			2.24	1.023
		1.2	1.87			2.24	0.731
		1.2	1.88			2.25	0.544
		1.25	1.92			2.39	4.55
Water	UCON 50HB400	0.79	1.16			0.91	0.11
		0.83	1.25			1.04	0.06
		0.83	1.27			1.06	0.06
		0.83	1.28			1.07	0.07
		0.83	1.29			1.07	0.07
		0.84	1.21			1.01	0.15
		0.84	1.24			1.03	0.08

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
		0.87	1.41			1.23	0.3
		0.88	1.31			1.16	0.15
		0.88	1.34			1.18	0.15
		0.88	1.37			1.2	0.2
		0.88	1.4			1.22	0.24
		0.89	1.26			1.12	0.26
		0.91	1.54			1.4	0.55
		0.92	1.52			1.4	0.74
		0.93	1.44			1.33	0.4
		0.93	1.48			1.38	0.57
		0.94	1.32			1.24	0.35
		0.94	1.4			1.31	0.35
		0.96	1.63			1.57	1.38
		0.96	1.64			1.57	0.03
		0.97	1.54			1.5	1.06
		0.97	1.59			1.55	1.54
		0.99	1.38			1.37	0.52
		0.99	1.49			1.46	0.7
		1	1.48			1.48	
		1	1.5			1.5	0
		1	1.5			1.5	0.04
		1.01	1.61			1.63	0.13
		1.01	1.61			1.63	0.13
		1.01	1.63			1.65	0
		1.01	1.69			1.71	0.4
		1.02	1.64			1.69	2.72
		1.02	1.66			1.7	2.11
		1.02	1.66			1.7	2.11
		1.02	1.66			1.7	2.11
		1.02	1.68			1.72	0.84
		1.02	1.69			1.72	1.5
		1.02	1.69			1.73	0.17
		1.02	1.69			1.73	0.29
		1.02	1.7			1.73	0.08
		1.02	1.71			1.74	0.44
		1.03	1.5			1.55	0.92
		1.03	1.5			1.55	0.92
		1.03	1.53			1.58	0.72
		1.03	1.53			1.58	0.72
		1.03	1.54			1.6	0.72
		1.03	1.54			1.6	0.72
		1.03	1.56			1.61	0.73
		1.03	1.56			1.61	0.73
		1.03	1.57			1.63	2.13
		1.03	1.57			1.61	0.92
		1.03	1.57			1.62	0.58

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
		1.03	1.57			1.63	2.13
		1.03	1.57			1.61	0.92
		1.03	1.57			1.62	0.58
		1.03	1.57			1.63	2.13
		1.03	1.58			1.63	1.25
		1.03	1.58			1.63	1.25
		1.03	1.6			1.65	0.47
		1.03	1.6			1.65	0.47
		1.03	1.61			1.65	0.56
		1.03	1.61			1.65	0.56
		1.03	1.62			1.67	0.38
		1.03	1.62			1.67	0.38
		1.03	1.63			1.68	0.9
		1.03	1.63			1.68	0.9
		1.03	1.64			1.69	2.72
		1.03	1.65			1.7	1.53
		1.03	1.65			1.7	0.46
		1.03	1.65			1.7	0.32
		1.03	1.65			1.7	1.53
		1.03	1.65			1.7	0.46
		1.03	1.65			1.7	0.32
		1.04	1.44			1.5	0.94
		1.04	1.44			1.5	0.94
		1.04	1.46			1.52	1.05
		1.04	1.48			1.53	1.02
		1.04	1.51			1.57	0.85
Water, D ₂ O		1.104	1.4			1.55	
Xylene hexafluoride		1.37	0.88			1.21	
Solids (Metals and Alloys)							
Aluminum		2.7	6.32	3.1			17.1
	Duraluminum	2.71	6.32	3.1			17.1
Al 1100-0	2S0	2.71	6.35	3.1	2.9		17.2
Al 2014	14S	2.8	6.32	3.1			17.7
Al 2024 T4	24ST	2.77	6.37	3.2	2.95		17.6
Al 2117 T4	17ST	2.8	6.5	3.1			18.2
Antimony	Sb		3.4				
Bearing Babbit		10.1	2.3				23.2
Beryllium		1.82	12.9	8.9	7.87		23.5
Bismuth		9.8	2.18	1.1			21.4
Brass	70% Cu-30% Zn	8.64	4.7	2.1			40.6
		8.56	4.28	2			36.6
	Half Hard	8.1	3.83	2.1			31.0
	Naval	8.42	4.43	2.1	1.95		37.3
Bronze	Phospho	8.86	3.53	2.2	2.01		31.3
Cadmium	Cd	8.6	2.8	1.5			42.0
		8.64	2.78	1.5			24.0

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Cesium		1.88	0.97				1.82
Columbium		8.57	4.92	2.1			42.2
Constantan		8.88	5.24	2.6			46.5
Copper		8.93	4.66	2.3	1.93		41.6
Copper, rolled	Cu	8.9	5.01	2.3			44.6
E-Solder		2.71	1.9	1			5.14
Gallium		5.95	2.74				16.3
Germanium		5.47	5.41				29.6
Gold	Hard drawn	19.32	3.24	1.2			62.6
Hafnium			3.84				
Inconel		8.25	5.72	3	2.79		64.5
Indium		7.3	2.22				16.2
Iron		7.7	5.9	3.2	2.79		45.4
	Cast	7.22	4.6	2.6			33.2
Lead		11.4	2.16	0.7	0.63		24.6
	5% Antimony	10.9	2.17	0.8	0.74		23.7
Magnesium		1.74	6.31				11.0
	AM-35	1.74	5.79	3.1	2.87		10.1
	FS-1	1.69	5.47	3			9.2
	J-1	1.7	5.67	3			9.6
	M	1.75	5.76	3.1			10.1
	O-1	1.82	5.8	3			10.6
	ZK-60A-TS	1.83	5.71	3.1			10.4
		1.72	5.8	3			10.0
Manganese		7.39	4.66	2.4			34.4
Molybdenum		10.2	6.29	3.4	3.11		64.2
Monel		8.83	6.02	2.7	1.96		53.2
Nickel		8.88	5.63	3	2.64		50.0
Nickel-silver		11.2	3.58	2.2			40.0
Platinum		21.4	3.96	1.7			84.7
Plutonium			1.79				28.2
	1% Gallium		1.82				28.6
Potassium		0.83	1.82				1.51
Rubidium		1.53	1.26				1.93
Silver		10.5	3.6	1.6			37.8
	Nickel	8.75	4.62	2.3	1.69		40.4
	Germanium	8.7	4.76				41.4
Steel	302 Cres	8.03	5.66	3.1	3.12		45.4
	347 Cres	7.91	5.74	3.1			45.4
	410 Cres	7.67	7.39	3	2.16		56.7
	1020	7.71	5.89	3.2			45.4
	1095	7.8	5.9	3.2			51.0
	4150	7.84	5.86	2.8			45.9
		7.82	5.89	3.2			46.1
		7.81	5.87	3.2			45.8
		7.8	5.82	2.8			45.4

Material	Comments	Density (g/cm ³)	V ₁ (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
	4340	7.8	5.85	3.2			51.0
	Mild	7.8	5.9	3.2			46.00
	Stainless 347	7.89	5.79	3.1			45.70
Tantalum		16.6	4.1	2.9			54.8
Thallium		11.9	1.62				19.3
Thorium		11.3	2.4	1.6			33.2
Tin		7.3	3.3	1.7			24.1
Titanium		4.5	6.07	3.1			27.3
		4.48	6.1	3.1			27.3
Tungsten		19.25	5.18	2.9	2.65		99.7
Uranium		18.5	3.4	2			63.0
Vanadium		6.03	6	2.8			36.2
Zinc		7.1	4.17	2.4			29.6
Zircalloy			4.72	2.4			44.2
Zirconium		6.48	4.65	2.3			30.1
Solids (Ceramics)							
Ammonium	502/ 118.9:1	1.35	2.73		3.69		
dihydrogen	502/ 118.5:1	1.35	2.67		3.60		
phosphate (ADP)			3.28				
Arsenic trisulfide		3.2	2.58	1.4			8.25
Barium titanate		5.55	5.64	2.9			33.5
Boron carbide		2.4	11				26.4
Brick		1.7	4.3				7.40
		3.6	3.65	2.6			15.3
Calcium fluoride	CaFl. X-cut		6.74				
Clay rock		2.5	3.48	3.4			14.2
Concrete		2.6	3.1				8.00
Flint		3.6	4.26	3			18.9
Glass	Crown	2.24	5.1	2.8			11.4
	205 Sheet	2.49	5.66				14.1
	FK3	2.26	4.91	2.9			11.1
	FK6	2.28	4.43	2.5			10.1
	Flint	3.6	4.5				16.0
	Macor	2.54	5.51				14.0
	Plate	2.75	5.71				10.7
	Pyrex	2.24	5.64	3.3			13.1
	Quartz	2.2	5.57	3.4			14.5
	Silica	2.2	5.9				13.0
	Soda lime	2.24	6				13.4
	T1K	2.38	4.38				10.5
	Window		6.79	3.4			
Glass crown	Reg.	2.6	5.66	3.5			14.5
Granite		4.1	6.5				26.8

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Graphite	Pyrolytic Pressed	1.46	4.6			6.60	
		1.8	2.4			4.10	
Hydrogen Ice	Solid at 4.2 K	0.089	2.19			0.19	
		0.92	3.6			3.20	
Ivory		2.65	3.99	3.3		16.4	
		2.17	3.01			10.4	
Leadmeta niobate	PbNbO ₃ K-81 K-83 K-85	6.2	3.3			20.5	
		6.2	3.3			20.5	
		4.3	5.33			22.9	
		5.5	3.35			18.4	
Lead zirconate titanate	PbZrTiO ₃	7.75	3.28			29.3	
		7.5	4			30.0	
		7.45	4.2			31.3	
		7.43	4.44			33.0	
		7.95	4.72			37.5	
Lithium niobate	46 Rot. Y-cut Z-cut Y-cut	4.7	7.08			33.0	
		4.64	7.33			34.0	
			6.88				
Lithium sulfate	Y-cut	2.06	5.46			11.2	
Marble		2.8	3.8			10.5	
Porcelain		2.3	5.9			13.50	
Potassium bromide			3.38				
Potassium chloride			4.14				
Potassium sodium niobate		4.46	6.94			31.0	
PZT-2		7.6	4.41	1.7		31.3	
PZT-4		7.5	4.6	1.9		34.5	
PZT-5A		7.75	4.35	1.7		33.7	
PZT-5H		7.5	4.56	1.8		34.2	
Quartz		6.82	5.66			15.2	
Salt	X-cut NaCl	2.65	5.75			15.3	
		2.17	4.85			10.5	
Salt, rochelle	KNaC ₄ H ₄ O ₆ X dir	2.2	5.36	3.8		13.1	
			2.47				
Salt, rock Sapphire	X dir		4.78				
		2.6	9.8			11.7	
Silica, fused	Al ₂ O ₃	3.98	11.2			44.5	
		2.2	5.96	3.8		13.1	
Silicon	Anisotropic	2.33	9			21.0	
Silicon carbide		13.8	6.66			91.8	
Silicon nitride		3.27	11	6.3		36.0	
Slate			4.5				
Sodium bismuth titanate		3	4.5			13.5	
		6.5	4.06			26.4	
Sodium bromide	NaBr		2.79				

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Sulfur			1.35				
Titanium carbide		5.15	8.27	5.2		42.6	
Tourmaline	Z-cut	3.1	7.54			23.4	
Uranium oxide	UO ₂		5.18			56.7	
Zinc oxide		5.68	6.4	3		36.4	
Solids (Polymer)							
ABS	Acrylonitrile	1.04	2.11		2.20		
Acrylic		1.2	2.7		3.24		
Acrylic resin		1.18	2.67	1.1		3.15	
Araldite	502/956	1.16	2.62		4.04		
Bakelite		1.4	2.59			3.63	
		1.9	1.9			4.80	
Butyl rubber		1.11	1.8			2.00	
Carbon, pyrolytic	Soft	2.21	3.31			7.31	
Carbon, vitreous		1.47	4.26	2.7		6.26	
Celcon	Acetal copolymer	1.41	2.51			3.54	
Cellulose acetate		1.3	2.45			3.19	
Cycolac	Acrylonitrile– butadiene–styrene		2.27			2.49	
ECHOGEL 1265	100PHA of B	9.19	1.32			12.2	
		1.4	1.7			2.38	
		1.1	1.71			1.90	
EPON 828	MPDA	1.21	2.83	1.2		3.40	
EPOTEK 301		1.08	2.64			2.85	
EPOTEK 330		1.14	2.57			2.94	
EPOTEK H70S		1.68	2.91			4.88	
EPOTEK V6	10PHA of B	1.23	2.55			3.14	
		1.23	2.61			3.21	
		1.26	2.55			3.22	
		1.25	2.6			3.25	
Epoxy	Silver	3.098	1.89			5.85	
		3.383	1.87			6.31	
EPX-1 or EPX-2	100PHA of B	1.1	2.44			2.68	
Ethyl vinyl acetate		0.94	1.8			1.69	
		0.95	1.68			1.60	
		0.93	1.86			1.72	
Delrin		1.36	2.47			3.36	
	Acetal homopolymer	1.42	2.52			3.57	
DER317	10.5PHR DEH20	1.18	2.75			3.25	
		2.23	2.07			4.61	
	13.5PHR MPDA	1.6	2.4			3.84	
		2.03	2.19			4.44	
		3.4	1.86	0.9		6.40	
	9PHR DEH20	7.27	1.5			10.9	
		2.23	2.03	1		4.53	
		2.37	1.93			4.58	

Material	Comments	Density (g/cm ³)	V ₁ (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
DER332	10PHR DEH20	1.76	3.18	1.6		5.58	
		1.2	2.6			3.11	
	10.5PHR DEH20	1.29	2.65			3.41	
		1.26	2.61			3.29	
		1.37	2.75			3.78	
	11PHR DEH20	1.72	2.35			4.05	
		1.29	2.71			3.49	
	14PHR MPDA	1.25	2.59			3.24	
	15PHR MPDA	1.54	2.78	1.5		4.27	
		1.49	2.8	1.4		4.18	
		1.24	2.66			3.30	
		1.24	2.55	1.2		3.16	
		2.15	3.75			8.06	
		2.24	3.9			8.74	
		6.45	1.75			11.3	
		64PHR V140	1.13	2.36			2.65
	75PHR V140	1.12	2.35			2.62	
	100PHR V140	1.1	2.32			2.55	
		1.13	2.27			2.55	
	Glucose		1.16	2.36			2.74
		1.56	3.2			5.00	
Hysol	C8-4143 / 3404	1.58	2.85			4.52	
		3.17	2.16			7.04	
	C9-4183/3561	2.14	2.49			5.33	
		1.8	2.62			4.70	
		1.48	2.92			4.30	
		2.66	2.3			6.10	
	C8-4412	1.68	2.02			3.39	
		1.5	2.32			3.49	
Hysol	R9-2039/3404						
Ivory			3.01				
Kel-F			1.79				
Kydex		1.35	2.22			2.99	
Lucite	Polymethylacrylate	1.29	2.72			3.50	
		1.18	2.68	1.3		3.16	
		1.15	2.7	1.1		3.10	
Marlex 5003	High-density polyethylene	0.95	2.56			2.43	
Melopas		1.7	2.9			4.93	
Micarta	Linen base		3				
Mylar		1.18	2.54			3.00	
Neoprene		1.31	1.6			2.10	
Noryl	Polyphenylene oxide	1.08	2.27			2.45	
Nylon 6-6		1.12	2.6	1.1		2.90	

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Penton	Chlorinated polyether Syntactic foam (33 lb/ft ³)	1.4	2.57			3.60	
		0.53	2.57			1.36	
Phenolic Plexiglas	UVA UVAII	1.34	1.42			1.90	
		1.27	2.76			3.51	
Polyamide Polycarbonate	Lexan	1.18	2.73	1.4		3.22	
			2.6			2.90	
Polyester	Casting Resin	1.18	2.3			2.71	
Polyethylene	Low density	1.07	2.29			2.86	
		0.92	2.06		1.90	22	26.5
Polyisobutylene	TCI HD. LB-861	1.1	2.67			2.80	
		0.96	1.6			2.33	
			2.43				
Polypropylene	mol. wt. 200 Profax 6423		1.49				
		0.901	1.85				
Polysulfone Polystyrene	Styron 666	0.88	2.49			2.24	
		1.24	2.74			2.40	
Polyurethane	RP-6400 RP-6401	1.24	2.24			2.78	
		1.1	2.67			2.80	
Polyvinyl chloride (PVC)	RP-6402 RP-6403 RP-6405 RP-6410 RP-6413 RP-6414 RP-6422 EN-9 REN plastic	1.05	2.4			2.52	
		1.04	1.5			1.56	
		1.07	1.71		1.83	35	73
		1.07	1.63			1.74	
		1.08	1.77			1.91	
		1.1	1.87			2.05	
		1.3	2.09			2.36	
		1.04	1.71		1.78	36	73
		1.04	1.33			1.38	
		1.04	1.71		1.78	21	35.2
		1.04	1.65			1.66	
		1.05	1.78			1.86	
		1.05	1.85		1.94	18	35.2
		1.04	1.6			1.66	
1.01	1.68			1.70			
Polyvinylbutyral Polyvinylidene difluoride Profax Refrasil	Polypropylene	1.07	1.71			1.83	35
		1.04	1.49			1.55	36
		1.04	1.62			1.69	15
		1.04	1.71			1.78	21
		1.05	1.85			1.92	18
		1.04	1.62		1.69	14	27.6
		1.45	2.27			3.31	
		1.11	2.35			2.60	
		1.79	2.3			4.20	
			2.79			2.51	
	3.75			6.49			

Material	Comments	Density (g/cm ³)	V _l (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Rubber	BFG#6063-19-71	0.97	1.53			1.56	
	BFG#35080						
	Hard	1.1	1.45			2.64	
	Rho-C	1	1.55			1.55	
	Soft	0.95	0.07			1.00	
Scotchcast	XR2535	1.49	2.48			3.70	
Scotchply XP241	Syntactic foam (42 lb/ft ³)	0.65	2.84			1.84	
	Syntactic foam (38 lb/ft ³)	0.61	2.81			1.71	
Scotchply	XP241	0.65	2.84			1.84	
	SP1002	1.94	3.25			6.24	
Scotch tape	2.5 mils thick	1.16	1.9			2.08	
Silicon rubber	Sylgard 170	1.38	0.97			1.34	
	Sylgard 182	1.05	1.03			1.07	
		1.12	1.03			1.15	
	Sylgard 184	1.03	1.03			1.04	
	RTV-11	1.18	1.05			1.24	
	RTV-21	1.31	1.01			1.32	
	RTV-30	1.45	0.97			1.41	
	RTV-41	1.31	1.01			1.32	
	RTV-60	1.47	0.96			1.41	
	RTV-77	1.33	1.02			1.36	
	RTV-90	1.5	0.96			1.44	
	RTV-112	1.05	0.94			0.99	
	RTV-511	1.18	1.11			1.31	
	RTV-116	1.1	1.02			1.12	
	RTV-118	1.04	1.03			1.07	
	RTV-577	1.35	1.08			1.46	
	RTV-560	1.42	1.03			1.46	
	RTV-602	1.02	1.16			1.18	
	RTV-615	1.02	1.08			1.10	
	RTV-616	1.22	1.06			1.29	
	RTV-630	1.24	1.05			1.30	
	PRC 1933-2	1.48	0.95			1.40	
Silly Putty		1	1			1.00	
Stycast	1251-40	1.67	2.9	1.5		4.83	
		1.63	2.95			4.82	
		1.57	2.88			4.53	
		1.5	2.77			4.16	
	1264	1.19	2.22			2.64	
	2741	1.17	2.29			2.68	
	CPC-41	1.01	1.52			1.54	
	CPC-39	1.06	1.53			1.63	
Styrene 50D	Polystyrene	1.04	2.33			2.43	
Styron	Modified polystyrene	1.03	2.24			2.31	
Surlyn	1555 Ionomer	0.95	1.91			1.81	

Material	Comments	Density (g/cm ³)	V ₁ (km/s)	V _s (km/s)	Impedance (MRayl)	Attn (dB/cm/MHz)	Attn (dB/cm at 5 MHz)
Tapox	Epoxy	1.11	2.48			2.76	
Techform	EA700	1.2	2.63			3.14	
Teflon		2.14	1.39			2.97	
		2.2	1.35			2.97	
TPX	DX845	0.83	2.22		1.84	4.2	5.8
Tracon	2135 D	1.03	2.45			1.52	
	2143 D	1.05	2.37			2.50	
	2162 D	1.19	2.02			2.41	
	3011	1.2	2.12			2.54	
	401 ST	1.62	2.97			4.82	
Uvex			2.11				
WR 106-1	Fluoro elastomer		0.87				
Zytel-101	Nylon-101	1.14	2.71			3.08	
Solids (Natural)							
Ash	Along fiber		4.67				
Beech	Along fiber		3.34				
Beef			1.55			1.68	
Brain			1.49			1.55	
Cork			0.5				
Douglas Fir	Cross grain		1.4				
	With grain		4.8				
Elm			1.4			0.798	
Human			1.47			1.58	
Kidney			1.54			1.62	
Liver			1.54			1.65	
Maple	Along fiber		4.11				
Oak			4.47			3.60	
Pine	Along fiber		3.32				
Poplar	Along fiber		4.28				
Spleen			1.5			1.60	
Sycamore	Along fiber		4.46				
Water		0.88	4	2		3.50	
Wood	Cork	0.24	0.5			0.12	
	Elm		4.1				
	Oak	0.72	4			1.57	
	Pine	0.45	3.5			1.57	

REFERENCES

1. *Metals Handbook*, 3rd ed., Vol. 17: *Nondestructive Evaluation and Quality Control*, ASM International, Metals Park, OH, 1989.
2. M. R. Mitchell and O. Buck (Eds.), *Cyclic Deformation, Fracture, and Nondestructive Evaluation of Advanced Materials*, American Society for Testing and Materials, Philadelphia, PA, 1992.
3. H. J. Shapuk (Ed.), *Annual Book of ASTM Standards: E-7, Nondestructive Testing*, American Society for Testing and Materials, West Conshohocken, PA, 1997.

4. R. E. Green, Jr. (Ed.), *Nondestructive Characterization of Materials*, Vol. 8: *International Symposium on Nondestructive Characterization of Materials*, Plenum, New York, 1998.
5. *British Journal of Nondestructive Testing*, published from JUL 1959 until MAR 1994.
6. American Society for Nondestructive Testing, <http://www.asnt.org>.
7. Center for Nondestructive Evaluation, <http://www.cnde.iastate.edu>.
8. Center for Quality Engineering & Failure Prevention, <http://www.cqe.nwu.edu>.
9. Nondestructive Evaluation System Reliability Assessment, <http://www.ihserc.com>.
10. Nondestructive Testing Information Analysis Center, <http://www.ntiac.com>.
11. W. Altergott and E. Henneke (Eds.), *Characterization of Advanced Materials*, Plenum, New York, 1990.
12. R. Halmshaw, *Nondestructive Testing Handbook*, Chapman & Hall, London, 1991.
13. R. A. Kline, *Nondestructive Characterization of Materials*, Technomic Publishing, Lancaster, PA, 1992.
14. P. K. Mallick, "Nondestructive Tests," in P. K. Mallick (Ed.), *Composites Engineering Handbook*, Marcel Dekker, New York, 1997.
15. W. McGonagle, *Nondestructive Testing*, Gordon Breach, New York, 1961.
16. C. O. Ruud et al. (Eds.), *Nondestructive Characterization of Materials*, Vols. I–IV, Plenum, New York, 1986.
17. R. S. Sharpe, *Research Techniques in Nondestructive Testing*, Academic, New York, 1984.
18. J. Summerscales, "Manufacturing Defects in Fibre-Reinforced Plastic Composites," *Insight*, **36**(12), 936–942, 1994.
19. D. O. Thompson and D. E. Chimenti (Eds.), *Review of Progress in Quantitative Nondestructive Evaluation*, Plenum, New York, 1982–2000.
20. J. Boogaard and G. M. van Dijk (Eds.), *Nondestructive Testing: Proceedings of the 12th World Conference on Nondestructive Testing*, Elsevier Science, New York, 1989.
21. D. E. Bray and R. K. Stanley, *Nondestructive Evaluation, A Tool for Design, Manufacturing, and Service*, McGraw-Hill, New York, 1989.
22. M. H. Geier, *Quality Handbook for Composite Materials*, Chapman & Hall, London, 1994.
23. Online Journal Publication Service, <http://ojps.aip.org>.
24. *Journal of Composite Materials*, 2004.
25. Electronic journals, <http://lib-www.lanl.gov/cgi-bin/ejrnlsrch.cgi>.
26. Elsevier Science, <http://www.elsevier.com/homepage/electserv.htm>.
27. *Japanese Journal of Nondestructive Inspection*, <http://sparc5.kid.ee.cit.nihon-u.ac.jp/homepage/Eng.html>.
28. *Journal of Nondestructive Evaluation*, Springer, <http://link.springer.com/journal/10921>
29. *Journal of Micromechanics and Microengineering*, 2004.
30. British Institute of Non-Destructive Testing, <http://www.bindt.org/>, 1999.
31. IFANT, International Foundation for the Advancement of Nondestructive Testing, <http://www.ifant.org>.
32. Japan JSNDI, <http://sparc5.kid.ee.cit.nihon-u.ac.jp/homepage/Eng.html>.
33. SPIE, <http://spie.org/>.
34. Institute of Electrical and Electronic Engineers, <http://www.ieee.org/>.
35. IEEE-ASME, *Journal of Microelectromechanical Systems*, Vol. 2000, 2004.
36. American Society of Mechanical Engineers, <http://www.asme.org/>.
37. Center for Nondestructive Evaluation, <http://www.cnde.com>.
38. Airport and Aircraft Safety Research & Development, <http://www.asp.tc.faa.gov>.
39. Fraunhofer IZFP, <http://www.fhg.de/english/profile/institute/izfp/index.html>.
40. Stasuk Testing & Inspection, <http://www.nde.net>.
41. AFRL electronic journals, <http://www.wrs.afrl.af.mil/infores/library/ejournals.htm>.
42. Link, Springer Verlag, <http://link.springer-ny.com/>.

43. IBM Intellectual Property Network, <http://www.patents.ibm.com>.
44. Lavender International NDT, <http://www.lavender-ndt.co.uk/>.
45. *Trends in NDE Science and Technology, Proceedings of the 14th World Conferences on Nondestructive Testing*, Brookfield VT, Ashgate Publishing, 1997.
46. J. L. Rose and A. A. Tseng (Eds.), *New Directions in Nondestructive Evaluation of Advanced Materials*, American Society of Mechanical Engineers, New York, 1988.
47. *Journal of Intelligent Material Systems and Structures*, <http://www.techpub.com>.
48. *Smart Structures*, <http://www.adaptive-ss.com/>.
49. *Smart Materials and Structures*, <http://www.adaptive-ss.com/>, 2001.
50. *Smart Structures—Harvard*, http://iti.acns.nwu.edu/clear/infr/imat_smart.html.
51. N. Tracy (Ed.), *Liquid Penetrant Testing*, 3rd ed., Vol. 2 of *Nondestructive Testing Handbook*, P. Moore (ed.), American Society for Nondestructive Testing, Columbus, OH, 1999.
52. R. A. Quinn, *Industrial Radiography—Theory and Practice*, Eastman Kodak, Rochester, NY, 1980.
53. H. Burger, *Neutron Radiography; Methods, Capabilities and Applications*, Elsevier Science, New York, 1965.
54. R. H. Bossi, F. A. Iddings, and G. C. Wheeler (Eds.), *Radiographic Testing*, 3rd ed., Volume in *Nondestructive Testing Handbook*, P. Moore (Ed.), American Society for Nondestructive Testing, Columbus, OH, 2002.
55. R. H. Fassbender and D. J. Hagemier, “Low-Kilovoltage Radiography of Composites,” *Mater. Eval.*, **41**(7), 381–838, 1983.
56. A. S. Birks and J. Green, *Ultrasonic Testing*, 2nd ed., Vol. 7 of *Nondestructive Testing Handbook*, P. Intire (Ed.), American Society for Nondestructive Testing, Columbus, OH, 1991.
57. E. A. Ash and E. G. S. Paige, *Rayleigh Wave Theory and Application*, *Springer Series on Wave Phenomena*, Vols. **1 and 2**, Springer-Verlag, Berlin, 1985.
58. I. A. Viktorov, *Rayleigh and Lamb Waves*, Plenum, New York, 1967.
59. J. Krautkramer and H. Krautkramer, *Ultrasonic Testing of Materials*, 3rd ed., Springer-Verlag, New York, 1983.
60. R. B. Jones and D. E. W. Stone, “Toward an Ultrasonic-Attenuation Technique to Measure Void Content in Carbon-Fibre Composites,” *Nondestructive Testing*, **9**(3), 71–79, 1976.
61. T. S. Jones, “Inspection of Composites Using the Automated Ultrasonic Scanning System (AUSS),” *Mater. Eval.*, **43**(5), 746–753, 1985.
62. D. K. Hsu and T. C. Patton, “Development of Ultrasonic Inspection for Adhesive Bonds in Aging Aircraft,” *Mater. Eval.*, **51**(12), 1390–1397, 1993.
63. R. N. Swamy and A. M. A. H. Ali, “Assessment of In Situ Concrete Strength by Various Non-Destructive Tests,” *NDT Int.*, **17**(3), 139–146, 1984.
64. E. P. Papadakis and G. B. Chapman II, “Modification of a Commercial Ultrasonic Bond Tester for Quantitative Measurements in Sheet-Molding Compound Lap Joints,” *Mater. Eval.*, **51**(4), 496–500, 1993.
65. D. J. Hagemier, “Bonded Joints and Nondestructive Testing—1,” *Nondestructive Testing*, **4**(12), 401–406, 1971.
66. D. J. Hagemier, “Bonded Joints and Nondestructive Testing—2,” *Nondestructive Testing*, **5**(2), 38–47, 1972.
67. D. J. Hagemier, “Nondestructive Testing of Bonded Metal-to-Metal Joints—2,” *Nondestructive Testing*, **5**(6), 144–153, 1972.
68. J. T. Schmidt and K. Skeie (Eds.), *Magnetic Particle Testing*, 2nd ed., Vol. 6 of *Nondestructive Testing Handbook*, P. McIntire (Ed.), American Society for Nondestructive Testing, Columbus, OH, 2001.
69. X. P. V. Maldague (Ed.), *Infrared and Thermal Testing*, 3rd ed., Vol. 3 of *Nondestructive Testing Handbook*, P. Moore (Ed.), American Society for Nondestructive Testing, Columbus, OH, 2001.
70. R. K. Stanley (Ed.), *Special Nondestructive Testing Methods*, 2nd ed., Vol. 9 of *Nondestructive Testing Handbook*, P. O. Moore and P. McIntire (Eds.), American Society for Nondestructive Testing, Columbus, OH, 1995.
71. T. B. Zorc (Ed.), *Nondestructive Evaluation and Quality Control*, 9th ed., Vol. 17 of *Metals Handbook*, ASM International, Metals Park, OH, 1989.

72. S. S. Udpa (Ed.), *Electromagnetic Testing*, 3rd ed., Vol. 5 of *Nondestructive Testing Handbook*, P. Moore (Ed.), American Society for Nondestructive Testing, Columbus, OH, 2004.
73. F. Förster, "Theoretische und experimentelle Grundlagen der zerstörungsfreien Werkstoffprüfung mit Wirbelstromverfahren, I. Das Tastpulverfahren," *Zeitschrift für Metallkunde*, **43**, 163–171, 1952.
74. S. N. Vernon, "Parametric Eddy Current Defect Depth Model and Its Application to Graphite Epoxy," *NDT Int.*, **22**(3), 139–148, 1989.
75. Wincheski, B. et al., "Development of Giant Magnetoresistive Inspection System for Detection of Deep Fatigue Cracks under Airframe Fasteners," in *Review of Progress in Quantitative Nondestructive Evaluation*, 2002.
76. J. Blitz, *Electrical and Magnetic Method of Non-Destructive Testing*, Chapman & Hall, London, 1997.
77. H. L. Libby, *Introduction to Electromagnetic Nondestructive Test Methods*, Wiley-Interscience, New York, 1971.
78. Wang, T. T., et al., "Effects of Bonding Defects on Shear Strength in Tension of Lap Shear Joints Having Brittle Adhesives," *J. Appl. Polymer Sci.*, **16**(8), 1901–1909, 1972.
79. Frock, B. G., et al., "Research on Advanced Methods for Aerospace Structures," Report No. UDR-TR-97-133, 1997, University of Dayton Research Institute, "Scans taken on 3 MAY 1996 show delams from LSP tests on Ti-6-4 turbine blade samples."
80. Crane, R. L., "NDE of Composites Materials," *Structural Integrity of Composite Materials and Structures: Optimisation of Micro-structural Design*, Isle of Capri Italy, 2001.
81. Crane, R. L., and Dillingham, G., "Composite Bond Inspection," *J. Mater. Sci.*, **43**(20), 6681–6694, 2008.
82. Dillingham, G., and Oakley, B., "Surface Energy and Adhesion in Composite–Composite Adhesive Bonds," *J. Adhesion*, **82**(4), 407–426, 2006.
83. Dillingham, G., et al., "Quantitative Detection of Peel Ply Derived Contaminants via Wettability Measurements," *J. Adhesion Sci. Tech.*, **26**(10-11), 1563–1571, 2012.
84. Dillingham, G., "Qualification of Surface Preparation Processes for Bonded Aircraft Repair," SAMPE 2013: Long Beach CA, Society for the Advancement of Material and Process Engineering, 2013.

CHAPTER 16

MATERIALS HANDLING SYSTEM DESIGN

Sunderesh S. Heragu and Banu Ekren
University of Louisville
Louisville, Kentucky

1 INTRODUCTION	497	3.6 Hoists, Cranes, and Jibs	505
2 TEN PRINCIPLES OF MATERIAL HANDLING	498	3.7 Warehouse Material Handling Devices	505
2.1 Planning	498	3.8 Autonomous Vehicle Storage and Retrieval System	505
2.2 Standardization	498	4 HOW TO CHOOSE THE “RIGHT” EQUIPMENT	506
2.3 Work	499	5 MULTIOBJECTIVE MODEL FOR OPERATION ALLOCATION AND SELECTION IN FMS DESIGN	506
2.4 Ergonomics	499	6 WAREHOUSING	509
2.5 Unit Load	500	6.1 Just-in-Time Manufacturing	509
2.6 Space Utilization	500	6.2 Warehouse Functions	509
2.7 System	501	6.3 Inverse Storage	510
2.8 Automation	502	7 AVS/RS CASE STUDY	510
2.9 Environment	503	REFERENCES	511
2.10 Life Cycle	503		
3 TYPES OF MATERIAL HANDLING EQUIPMENT	504		
3.1 Conveyors	504		
3.2 Palletizers	504		
3.3 Trucks	504		
3.4 Robots	505		
3.5 Automated Guided Vehicles	505		

1 INTRODUCTION*

Material handling systems consist of discrete or continuous resources to move entities from one location to another. They are more common in manufacturing systems compared to service systems. Material movement occurs everywhere in a factory or warehouse—before, during, and after processing. Apple¹ notes that material handling can account for up to 80% of production activity. Although material movement does not add value in the manufacturing process, half of the company’s operation costs are material handling costs.² Therefore, keeping the material handling activity at a minimum is very important for companies.

Due to the increasing demand for a high variety of products and shorter response times in today’s manufacturing industry, there is a need for highly flexible and efficient material handling systems. In the design of a material handling system, facility layout, product routings, and material flow control must be considered. In addition, various other factors must be

*Many of the sections in this chapter have been reproduced from Chapter 11 of Ref. 3 with permission.

considered in an integrated manner. The next section describes the 10 principles of material handling as developed by the Material Handling Industry of America (MHIA). It presents a guideline for selecting equipment, designing a layout, and standardizing, managing, and controlling the material movement as well as the handling system. Another section describes the common types of material handling systems. This chapter also discusses types of equipment, how to select material handling equipment, an operating model for material handling, and warehousing issues. It ends with a case study that implements some of these issues.

2 TEN PRINCIPLES OF MATERIAL HANDLING

If material handling is designed properly, it provides an important support to the production process. Following is a list of 10 principles as developed by the MHIA, which can be used as a guide for designing material handling systems.

2.1 Planning

A *plan* is a prescribed course of action that is defined in advance of implementation. In its simplest form, a material handling plan defines the material (what) and the moves (when and where); together, they define the method (how and who). Five key aspects must be considered in developing a plan:

1. The plan should be developed in consultation between the planner(s) and all who will use and benefit from the equipment to be employed.
2. Success in planning large-scale material handling projects generally requires a team approach involving suppliers, consultants when appropriate, and end-user specialists from management, engineering, computer and information systems, finance, and operations.
3. The material handling plan should reflect the strategic objectives of the organization as well as the more immediate needs.
4. The plan should document existing methods and problems, physical and economic constraints, and future requirements and goals.
5. The plan should promote concurrent engineering of product, process design, process layout, and material handling methods, as opposed to independent and sequential design practices.

2.2 Standardization

Material handling methods, equipment, controls, and software should be standardized within the limits of achieving overall performance objectives and without sacrificing needed flexibility, modularity, and throughput. Standardization means less variety and customization in the methods and equipment employed. There are three key aspects of achieving standardization:

1. The planner should select methods and equipment that can perform a variety of tasks under a variety of operating conditions and in anticipation of changing future requirements.
2. Standardization applies to sizes of containers and other load-forming components as well as operating procedures and equipment.
3. Standardization, flexibility, and modularity must not be incompatible.

2.3 Work

The measure of work is material handling flow (volume, weight, or count per unit of time) multiplied by the distance moved. Material handling work should be minimized without sacrificing productivity or the level of service required of the operation. Five key points are important in optimizing the work:

1. Simplifying processes by reducing, combining, shortening, or eliminating unnecessary moves will reduce work.
2. Consider each pickup and set-down—that is, placing material in and out of storage—as distinct moves and components of the distance moved.
3. Process methods, operation sequences, and process/equipment layouts should be prepared that support the work minimization objective.
4. Where possible, gravity should be used to move materials or to assist in their movement while respecting consideration of safety and the potential for product damage (see Fig. 1).
5. The shortest distance between two points is a straight line.

2.4 Ergonomics

Ergonomics is the science that seeks to adapt work or working conditions to suit the abilities of the worker. Human capabilities and limitations must be recognized and respected in the design of material handling tasks and equipment to ensure safe and effective operations. There are two key points in the ergonomic principles:

1. Equipment should be selected that eliminates repetitive and strenuous manual labor and that effectively interacts with human operators and users. The ergonomic principle embraces both physical and mental tasks.
2. The material handling workplace and the equipment employed to assist in that work must be designed so they are safe for people.



Figure 1 Gravity roller conveyor. Courtesy of Pentek.

2.5 Unit Load

A unit load is one that can be stored or moved as a single entity at one time, such as a pallet, container, or tote, regardless of the number of individual items that make up the load. Unit loads shall be appropriately sized and configured in a way that achieves the material flow and inventory objectives at each stage in the supply chain. When unit load is used in material flow, six key aspects deserve attention:

1. Less effort and work are required to collect and move many individual items as a single load than to move many items one at a time.
2. Load size and composition may change as material and products move through stages of manufacturing and the resulting distribution channels.
3. Large unit loads are common both pre- and postmanufacturing in the form of raw materials and finished goods.
4. During manufacturing, smaller unit loads, including as few as one item, yield less in-process inventory and shorter item throughput times.
5. Smaller unit loads are consistent with manufacturing strategies that embrace operating objectives such as flexibility, continuous flow, and just-in-time delivery.
6. Unit loads composed of a mix of different items are consistent with just-in-time and/or customized supply strategies as long as item selectivity is not compromised.

2.6 Space Utilization

Space in material handling is three dimensional and therefore is counted as cubic space. Effective and efficient use must be made of all available space. This is a three-step process:

1. Eliminate cluttered and unorganized spaces and blocked aisles in work areas (see Fig. 2).
2. In storage areas, balance the objective of maximizing storage density against accessibility and selectivity. If items are going to be in the warehouse for a long time, storage



Figure 2 Retrieving material in blocked aisles.

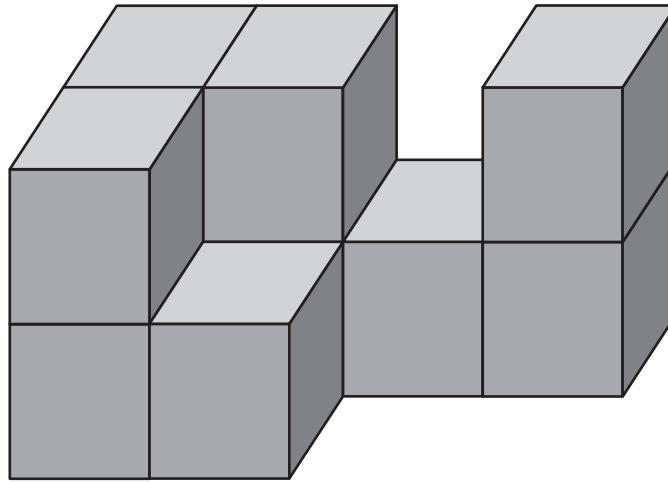


Figure 3 Honeycombing loss.

density is an important consideration. Avoid honeycombing loss (Fig. 3). If items enter and leave the warehouse frequently, their accessibility and selectivity are important. If the storage density is too high to access or select the stored product, high storage density may not be beneficial.

3. Consider the use of overhead space when transporting loads within a facility. Cube per order index (COI) storage policy is often used in a warehouse. COI is a storage policy in which each item is allocated warehouse space based on the ratio of its storage space requirements (its cube) to the number of storage/retrieval transactions for that item. Items are listed in a nondecreasing order of their COI ratios. The first item in the list is allocated to the required number of storage spaces that are closest to the input/output (I/O) point; the second item is allocated to the required number of storage spaces that are next closest to the I/O point, and so on. Figure 4 shows an interactive *playspace* in the “ten principles of materials handling” CD that allows a learner to understand the fundamental concepts of the COI policy.

2.7 System

A *system* is a collection of interacting or interdependent entities that form a unified whole. Material movement and storage activities should be fully integrated to form a coordinated operational system that spans receiving, inspection, storage, production, assembly, packaging, unitizing, order selection, shipping, transportation, and the handling of returns. Here are five key aspects of the system principle:

1. Systems integration should encompass the entire supply chain, including reverse logistics. It should include suppliers, manufacturers, distributors, and customers.
2. Inventory levels should be minimized at all stages of production and distribution while respecting considerations of process variability and customer service.
3. Information flow and physical material flow should be integrated and treated as concurrent activities.
4. Methods should be provided for easily identifying materials and products, for determining their location and status within facilities and within the supply chain, and for

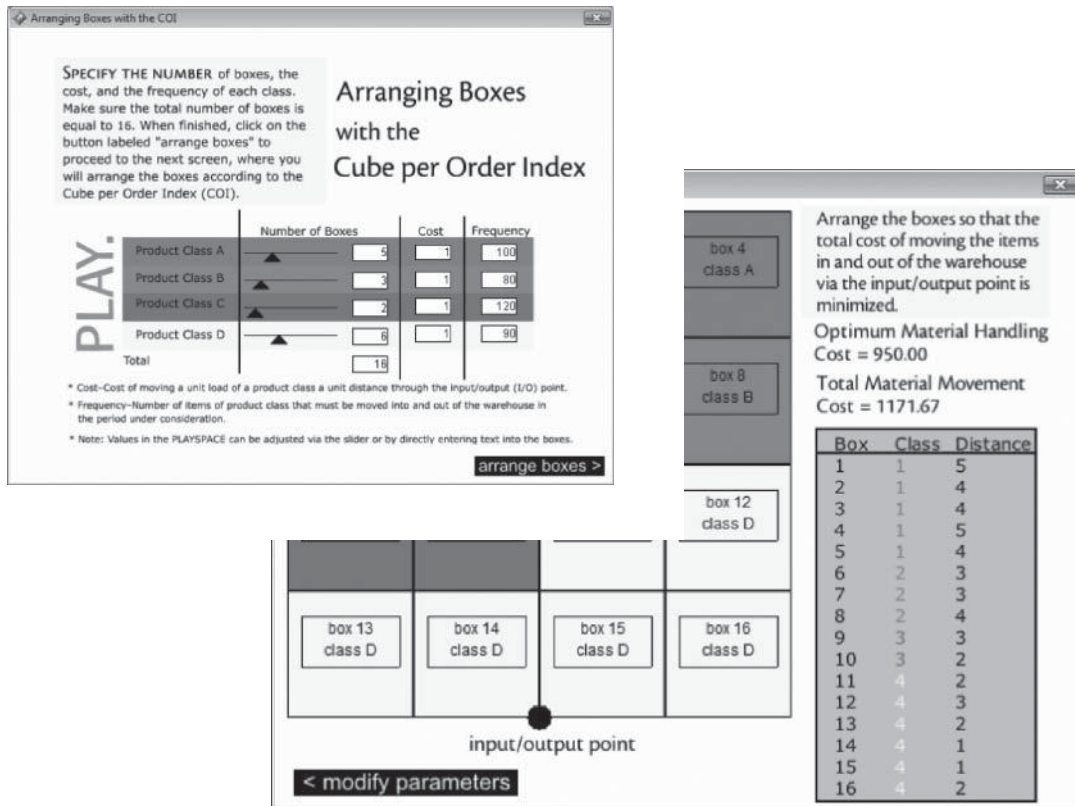


Figure 4 Example of COI policy.

controlling their movement. For instance, bar coding is the traditional method used for product identification. Radio frequency identification (RFID) uses radio waves to automatically identify objects as they move through the supply chain. The big difference between the two automatic data capture technologies is that bar coding is a line-of-sight technology. In other words, a scanner has to "see" the bar code to read it, which means people usually have to orient the bar code toward a scanner for it to be read. RFID tags can be read as long as they are within the range of a reader, even if there is no line of sight. Bar codes have other shortcomings as well. If a label is ripped, soiled, or falls off, there is no way to scan the item. Also, standard bar codes identify only the manufacturer and product, not the unique item. The bar code on one gallon of 2% milk is the same as on every other gallon of the same brand, making it impossible to identify which one might pass its expiration date first. RFID can identify items individually.

5. Customer requirements and expectations regarding quantity, quality, and on-time delivery should be met without exception.

2.8 Automation

Automation is concerned with the application of electromechanical devices, electronics, and computer-based systems to operate and control production and service activities. It suggests

the linking of multiple mechanical operations to create a system that can be controlled by programmed instructions. Material handling operations should be mechanized and/or automated where feasible to improve operational efficiency, increase responsiveness, improve consistency and predictability, decrease operating costs, and eliminate repetitive or potentially unsafe manual labor. There are four key points in automation:

1. Preexisting processes and methods should be simplified and/or reengineered before any efforts at installing mechanized or automated systems.
2. Computerized material handling systems should be considered where appropriate for effective integration of material flow and information management.
3. All items expected to be handled automatically must have features that accommodate mechanized and automated handling.
4. All interface issues should be treated as critical to successful automation, including equipment to equipment, equipment to load, equipment to operator, and control communications.

2.9 Environment

Environmental consciousness stems from a desire not to waste natural resources and to predict and eliminate the possible negative effects of our daily actions on the environment. Environmental impact and energy consumption should be considered as criteria when designing or selecting alternative equipment and material handling systems. Here are the three key points:

1. Containers, pallets, and other products used to form and protect unit loads should be designed for reusability when possible and/or biodegradability as appropriate.
2. Systems design should accommodate the handling of spent dunnage, empty containers, and other byproducts of material handling.
3. Materials specified as hazardous have special needs with regard to spill protection, combustibility, and other risks.

2.10 Life Cycle

Life-cycle costs include all cash flows that will occur between the time the first dollar is spent to plan or procure a new piece of equipment, or to put in place a new method, until that method and/or equipment is totally replaced. A thorough economic analysis should account for the entire life cycle of all material handling equipment and resulting systems. There are four key aspects:

1. Life-cycle costs include capital investment, installation, setup and equipment programming, training, system testing and acceptance, operating (labor, utilities, etc.), maintenance and repair, reuse value, and ultimate disposal.
2. A plan for preventive and predictive maintenance should be prepared for the equipment, and the estimated cost of maintenance and spare parts should be included in the economic analysis.
3. A long-range plan for replacement of the equipment when it becomes obsolete should be prepared.
4. Although measurable cost is a primary factor, it is certainly not the only factor in selecting among alternatives. Other factors of a strategic nature to the organization that form the basis for competition in the marketplace should be considered and quantified whenever possible.

These 10 principles are vital to material handling system design and operation. Most are qualitative in nature and require the industrial engineer to employ these principles when designing, analyzing, and operating material handling systems.

3 TYPES OF MATERIAL HANDLING EQUIPMENT

In this section, we list various equipments that actually transfer materials between the multiple stages of processing. There are a number of different types of material handling devices (MHDs), most of which move materials via material handling paths on the shop floor. However, there are some MHDs—such as cranes, hoists, and overhead conveyors—that utilize the space above the machines. The choice of a specific MHD depends on a number of factors, including cost, weight, size, and volume of the loads; space availability; and types of workstations. So, in some cases the materials handling system (MHS) interacts with the other subsystems. If we isolate MHS from other subsystems, we might get an optimal solution relative to the MHDs but one that is suboptimal for the entire system.

There are seven basic types of MHDs³: conveyors, palletizers, trucks, robots, automated guided vehicles (AGVs), hoists cranes and jibs, and warehouse material handling devices. In this section, we will introduce the seven basic types of MHDs. In the following section, we will discuss how to choose the “right” equipment and how to operate equipment in the “right” way.

3.1 Conveyors

Conveyors are fixed-path MHDs. In other words, conveyors should be considered only when the volume of parts or material to be transported is large and when the transported material is relatively uniform in size and shape. Depending on the application, there are many types of conveyors —accumulation conveyor, belt conveyor, bucket conveyor, can conveyor, chain conveyor, chute conveyor, gravity conveyor, power and free conveyor, pneumatic or vacuum conveyor, roller conveyor, screw conveyor, slat conveyor, tow line conveyor, trolley conveyor, and wheel conveyor. Our list is not meant to be complete, and other variations are possible. For example, belt conveyors may be classified as troughed belt conveyors (used for transporting bulky material such as coal) and magnetic belt conveyors (used for moving ferrous material against gravitational force). For the latest product information on conveyors and other types of material handling equipment, we strongly encourage the reader to refer to recent issues of *Material Handling Engineering* and *Modern Materials Handling*. These publications not only have articles illustrating use of the material handling equipment but also numerous product advertisements.

3.2 Palletizers

Palletizers are high-speed automated equipment used to palletize containers coming off production or assembly lines. With operator-friendly touch-screen controls, they palletize at the rate of a hundred cases per minute, palletize two lines of cases simultaneously, or simultaneously handle multiple products.

3.3 Trucks

Trucks are particularly useful when the material moved varies frequently in size, shape, and weight, when the volume of the parts or material moved is low, and when the number of trips required for each part is relatively small. There are several trucks in the market with different weight, cost, functionality, and other features. Hand truck, fork lift truck, pallet truck, platform truck, counterbalanced truck, tractor-trailer truck, and AGVs are some examples of trucks.

3.4 Robots

Robots are programmable devices that resemble the human arm. They are also capable of moving like the human arm and can perform functions such as weld, pick and place, and load and unload. Some advantages of using a robot are that they can perform complex repetitive tasks automatically and they can work in hazardous and uncomfortable environments that a human operator cannot work. The disadvantage is that robots are relatively expensive.

3.5 Automated Guided Vehicles

AGVs have become very popular, especially in the past decade, and will continue to be the dominant type of MHD in the years to come. The first system was installed in 1953, and the technology continues to expand. AGVs can be regarded as a type of specially designed robots. Their paths can be controlled in a number of different ways. They can be fully automated or semiautomated. AGVs are becoming more flexible with a wider range of applications using more diverse vehicle types, load transfer techniques, guide path arrangements, controls, and control interfaces. They can also be embedded into other MHDs.

3.6 Hoists, Cranes, and Jibs

These MHDs are preferred when the parts to be moved are bulky and require more space for transportation. Because the space above the machines is typically utilized only for carrying power and coolant lines, there is abundant room to transport bulky material. The movement of material in the overhead space does not affect production process and worker in a factory. The disadvantages of these MHDs are that they are expensive and time-consuming to install.

3.7 Warehouse Material Handling Devices

These are typically referred to as storage and retrieval systems. If they are automated to a high degree, they are referred to as automated storage and retrieval systems (AS/RSs). The primary functions of warehouse material handling devices are to store and retrieve materials as well as transport them between the pick/deposit (P/D) stations and the storage locations of the materials.

AS/RSs are capital-intensive systems. However, they offer a number of advantages, such as low labor and energy costs, high land or space utilization, high reliability and accuracy, and high throughput rates.

3.8 Autonomous Vehicle Storage and Retrieval System

Autonomous vehicle storage and retrieval systems (AVS/RSs) represent a relatively new technology for automated unit load storage systems. In this system, the autonomous vehicles function as storage/retrieval (S/R) devices. Within the storage rack, the key distinction of AVS/RSs relative to traditional crane-based AS/RSs is the movement patterns of the S/R device. In AS/RSs, aisle-captive storage cranes can move in the horizontal and vertical dimensions simultaneously to store or retrieve unit loads. In an AVS/RS, vehicles use a fixed number of lifts for vertical movement and follow rectilinear flow patterns for horizontal travel. Although the travel patterns in an AS/RS are generally more efficient within storage racks, an AVS/RS has a significant potential advantage in the adaptability of system throughput capacity to transactions demand by changing the number of vehicles operating in a fixed storage configuration. For example, decreasing the number of vehicles increases the transaction cycle times and utilization, which are also key measures of system performance.

4 HOW TO CHOOSE THE “RIGHT” EQUIPMENT

Apple¹ has suggested the use of the “material handling equation” in arriving at a material handling solution. The methodology uses six major questions: why (select material handling equipment), what (is the material to be moved), where and when (is the move to be made), how (will the move be made), and who (will make the move). All these six questions are extremely important and should be answered satisfactorily.

The material handling equation can be specified as: *Material* + *Move* = *Method*. Very often, when the *material* and *move* aspects are analyzed thoroughly, it automatically uncovers the appropriate material handling *method*. For example, analysis of the type and characteristics of *material* may reveal that the material is a large unit load on wooden pallets. Further analysis of the logistics, characteristics, and type of *move* may indicate that 6 m load/unload lift is required, distance traveled is 50 m, and some maneuvering is required while transporting the unit load. This suggests that a fork lift truck would be a suitable material handling device. Even further analysis of the method may tell us more about the specific features of the fork lift truck, for example, a narrow-aisle fork lift truck with a floor load capacity of $1/2$ ton.

5 MULTIOBJECTIVE MODEL FOR OPERATION ALLOCATION AND SELECTION IN FMS DESIGN

From both a conceptual and a computational point, only a few mathematical programming models have been proposed for the material handling system selection problem. Most of the studies have focused on material handling equipment optimization, rather than the entire material handling system. Sujono and Lashkari⁴ proposed a multiobjective model for selecting MHDs and allocating material handling transactions to them in flexible manufacturing system (FMS) design. They propose a model that integrates operation allocation (OA) and MHD selection problem. Their study is an extension of the Paulo et al.⁵ and Lashkari et al.⁶ studies. The main differences from the previous models are the new definition of the variables and the introduction of a new variable that links the selection of a machine to perform manufacturing operation with the material handling requirements of that operation. In addition, they include all the costs associated with material handling operations and suboperations, and the complete restructuring of the constraints that control the selection of the material handling equipment and their loading, in the objective function. Their model is presented as follows:

$h \in \{1, 2, \dots, H\}$	Major MH operations
$\hat{h} \in \{1, 2, \dots, \hat{H}\}$	MH suboperations
$e \in E_{j\hat{h}}\{1, 2, \dots, E\}$	Set of MH equipment that can handle the combination of MH operation/suboperation at machine j
$j \in J_{ips}\{1, 2, \dots, m\}$	Set of machines that can perform operation s of part type i under process plan p
PARAMETERS	
b_j	Time available on machine j
OC_{ipj}	Cost of performing operation s of part type i under process plan p on machine j (\$)
d_i	Demand for part type i (units)
SC_j	Setup cost of machine j (\$)
t_{ijp}	Time for performing operation s of part type i under process plan p on machine j

$T_{ijh\hat{h}e}$	MH cost of performing the combination of MH operation/suboperation for part type i on machine j using MH equipment e (\$)
L_e	Time available on MH equipment e
$I_{h\hat{h}e}$	Time for MH equipment e to perform the combination of MH operation/suboperation
\hat{W}_{it}	Relative weight of the product variable t on part type i
W_{et}	Relative weight of the product variable t on MH equipment e
$W_{h\hat{h}e}$	Relative degree of capability of MH equipment e to perform the combination of MH operation/suboperation
C_{ei}	Compatibility between MH equipment e and part type i

DECISION VARIABLES

$Z(ip) \in \{1, 0\}$	1 if part type i uses process plan p ; 0 otherwise
$Y_{sj(ip)} \in \{1, 0\}$	1 if machine j performs operation s of part type i under process plan p ; 0 otherwise
$A_{iiph\hat{h}} \in \{1, 0\}$	1 if part type i under process plan p requires the combination of MH operation/suboperation at machine j ; 0 otherwise
$X_{ijph\hat{h}e} \in \{1, 0\}$	1 if the combination of MH operation/suboperation requires MH equipment e at machine j where operation s of part type i under process plan p is performed; 0 otherwise
$M_j \in \{1, 0\}$	1 if machine j is selected; 0 otherwise
$De \in \{1, 0\}$	1 if MH equipment e is selected; 0 otherwise

The first part of the objective function is presented as

$$\begin{aligned}
 F_1 = & \sum_{i=1}^n d_i \sum_{p=1}^{p(i)} \sum_{s=1}^{S(i)p} \sum_{j \in J_{ips}} OC_{ipj} Y_{sj}(ip) + \sum_{j=1}^m SC_j M_j \\
 & + \sum_{i=1}^n d_i \sum_{p=1}^{p(i)} \sum_{s=1}^{S(i)p} \sum_{j \in J_{ips}} \sum_{h=1}^H \sum_{\hat{h}=1}^{\hat{H}} \sum_{e \in E_{jh}} T_{ijh\hat{h}e} X_{ijph\hat{h}e} \quad (1)
 \end{aligned}$$

The second part of the objective function is formulated as

$$F_2 = \sum_{e=1}^E \sum_{h=1}^H \sum_{\hat{h}=1}^{\hat{H}} W_{h\hat{h}e} \sum_{i=1}^n C_{ei} \sum_{p=1}^{p(i)} \sum_{s=1}^{S(ip)} \sum_{j \in J_{ips}} X_{ijph\hat{h}e} \quad (2)$$

where

$$C_{ei} = 1 - \frac{\sum_{t=1}^T |W_{et} - \hat{W}_{it}|}{4T}$$

Here, $T = 5$ and refers to the five major variables used to identify the dimensions of the characteristics mentioned by Ayres.⁷ Integer numbers are used to assign values to the subjective factors, W parameters, W_{et} , $W_{h\hat{h}e}$, and \hat{W}_{it} . The rating scales range from 0 to 5 for W_{et} and $W_{h\hat{h}e}$ and from 1 to 5 for \hat{W}_{it} .⁷ A 5 for W_{et} means that the piece of equipment is best suited to handle parts with a very high rating of product variable t . A 0 means do not allow this piece of equipment to handle parts with product variable t . A 5 for $\hat{W}_{h\hat{h}e}$ means that it is excellent in performing the operation/suboperation combination. And a 0 means that it is incapable of performing the

operation/suboperation combination. A 5 for W_{it} means that the part type exhibits a very high level of the key product variable t . And a 0 means that the part type exhibits a very low level of the key product variable t .

The first part of objective function's three terms indicates the manufacturing operation costs, the machine setup costs, and the MH operation costs, respectively. The second part of the objective function computes the overall compatibility of the MH equipment. As a result, the formulation of the problem is a multiobjective model seeking to strike a balance between the two objectives.

There are nine constraints in this model:

1. Each part type can use only one process plan:

$$\sum_{p=1}^{p(i)} Z(ip) = 1 \quad \forall i \quad (3)$$

2. For a given part type i under process plan p , each operation of the selected process plan is assigned to only one of the available machines:

$$\sum_{j \in J_{ips}} Y_{sj}(ip) = Z(ip) \quad \forall i, p, s \quad (4)$$

3. Once a machine is selected for operation s of part type i under process plan p , then all the $(h\hat{h})$ combinations corresponding to (sj) must be performed:

$$Y_{sj}(ip) = A_{sjh\hat{h}}(ip) \quad \forall i, p, s, j, h, \hat{h} \quad (5)$$

4. Each $h\hat{h}$ combination can be assigned to only the piece of available and capable MH equipment:

$$\sum_{e \in E_{jhh}} X_{ijph\hat{h}e} = A_{ijph\hat{h}} \quad \forall i, p, s, j, h, \hat{h} \quad (6)$$

5. At least one operation must be allocated to a selected machine:

$$\sum_{i=1}^n \sum_{p=1}^{P(i)} \sum_{s=1}^{S(ip)} \sum_{h=1}^H \sum_{\hat{h}=1}^{\hat{H}} \sum_{e \in E_{jhh}} X_{ijph\hat{h}e} \geq M_j \quad \forall j \quad (7)$$

6. The allocated operations cannot exceed the corresponding machine's capacity:

$$\sum_{i=1}^n d_i \sum_{p=1}^{P(i)} \sum_{s=1}^{S(ip)} \sum_{h=1}^H \sum_{\hat{h}=1}^{\hat{H}} \sum_{e \in E_{jhh}} t_{sj}(ip) X_{ijph\hat{h}e} \geq b_j M_j \quad \forall j \quad (8)$$

7. A specific MH equipment can be selected only if the corresponding *type* of equipment is selected:

$$D_e \leq D_{\hat{e}} \quad \forall e, \hat{e} \quad (9)$$

8. Each MH equipment selected must perform at least one operation:

$$\sum_{i=1}^n \sum_{h=1}^H \sum_{\hat{h}=1}^{\hat{H}} \sum_{p=1}^{P(i)} \sum_{s=1}^{S(ip)} \sum_{j \in J_{ips}} X_{ijph\hat{h}e} \geq D_e \quad \forall e \quad (10)$$

9. The MH equipment capacity cannot be exceeded:

$$\sum_{i=1}^n d_i \sum_{h=1}^H \sum_{\hat{h}=1}^{\hat{H}} \sum_{p=1}^{P(i)} \sum_{s=1}^{S(ip)} \sum_{j \in J_{ips}} X_{ijph\hat{h}e} \geq D_e \quad \forall e \quad (11)$$

6 WAREHOUSING

Many manufacturing and distribution companies maintain large warehouses to store in-process inventories or components received from an external supplier. They are involved in various stages of the sourcing, production, and distribution of goods, from raw materials through the finished goods. The true value of warehousing lies in having the right product in the right place at the right time. Thus, warehousing provides the time-and-place utility necessary for a company and is often one of the most costly elements. Therefore, its successful management is critical.

6.1 Just-in-Time Manufacturing

It has been argued that warehousing is a time-consuming and non-value-adding activity. Because additional paperwork and time are required to store items in storage spaces and retrieve them later when needed, the just-in-time (JIT) manufacturing philosophy suggests that one should do away with any kind of temporary storage and maintain a pull strategy in which items are produced only as and when they are required. That is, they should be produced at a certain stage of manufacturing only if they are required at the next stage.

JIT philosophy requires that the same approach be taken toward components received from suppliers. The supplier is considered as another (previous) stage in manufacturing. However, in practice, because the demand is continuous, that means that goods need to be always pulled through the supply chain to respond to demand quickly. The handling of returned goods is becoming increasingly important (e.g., Internet shopping may increase the handling of returned goods), and due to the uncertainty inherent in the supply chain, it is not possible to completely do away with temporary storage.

6.2 Warehouse Functions

Every warehouse should be designed to meet the specific requirements of the supply chain of which it is a part. In many cases, the need to provide better service to customers and be responsive to their needs appears to be the primary reason. Nevertheless, there are certain operations that are common to most warehouses:

- *Temporarily store goods.* To achieve economies of scale in production, transportation, and handling of goods, it is often necessary to store goods in warehouses and release them to customers as and when the demand occurs.
- *Put together customer orders.* Goods are received from order picking stock in the required quantities and at the required time to the warehouse to meet customer orders. For example, goods can be received from suppliers as whole pallet quantities but are ordered by customers in less than pallet quantities.
- *Serve as a customer service facility.* In some cases, warehouses ship goods to customers and therefore are in direct contact with them. So, a warehouse can serve as a customer service facility and handle replacement of damaged or faulty goods, conduct market surveys, and even provide after sales service. For example, many Korean electronic goods manufacturers let warehouses handle repair and do after sales service in North America.
- *Protect goods.* Sometimes manufactured goods are stored in warehouses to protect them against theft, fire, floods, and weather elements because warehouses are generally secure and well equipped.
- *Segregate hazardous or contaminated materials.* Safety codes may not allow storage of hazardous materials near the manufacturing plant. Because no manufacturing takes

place in a warehouse, this may be an ideal place to segregate and store hazardous and contaminated materials.

- *Perform value-added services.* In many warehouses after picking, goods are brought together and consolidated as completed orders ready to be dispatched to customers. This can involve packing into dispatch outer cases and cartons, and stretch and shrink wrapping for load protection and stability, inspecting, and testing. Here, inspection and testing do not add value to the product. However, we have included them because they may be a necessary function because of company policy or federal regulations.
- *Store seasonal inventory.* It is always difficult to forecast product demand accurately in many businesses. Therefore, it may be important to carry inventory and safety stocks to meet unexpected surges in demand. Some companies that produce seasonal products—for example, lawn mowers and snow throwers—may have excess inventory left over at the end of the season and have to store the unsold items in a warehouse.

A typical warehouse consists of two main elements:

1. Storage medium
2. Material handling system

In addition, there is a building that encloses the storage medium, goods, and the S/R system. Because the main purpose of the building is to protect its contents from theft and weather elements, it is made of strong, lightweight material. So, warehouses come in different shapes, sizes, and heights, depending on a number of factors, including the kind of goods stored inside, volume, and type of S/R systems used. For example, the Nike warehouse in Laakdal, Belgium, covers a total area of 1 million square feet. Its high-bay storage is almost 100 feet in height, occupies roughly half of the total warehouse space, and is served by 26 man-aboard stacker cranes.

6.3 Inverse Storage

There is limited landfill space available for dumping wastes created throughout the supply chain. And the increasing cost of landfills, environmental laws and regulation, and economic viability of environmental strategies are pushing manufacturers nowadays to consider reverse supply chain—also known as *reverse logistics*—management.

Manufacturers now must take full responsibility for their products through the product's life cycle or they may be subject to legal action. For example, new laws regarding the disposal of motor or engine oil, vehicle batteries, and tires place the disposal responsibility on the manufacturer once these products have passed their useful life. Many manufacturers also realize that reverse logistics offers the opportunity to recycle and reuse product components and reduce the cost and the amount of waste. Therefore, manufacturers are developing disposition stocking areas and collecting used or expired original products from the customer and reshipping to their stocking places. For example, Kodak's single-use camera has a remarkable success story involving the inverse logistics philosophy. The products are collected in a stocking place to be remanufactured. In the United States, 63% return rate has been achieved for recycling. The details about the procedure can be obtained from Kodak's website (<http://www.kodak.com/US/en/corp/environment/performance/recycling.html>).

7 AVS/RS CASE STUDY

Savoye Logistics is a European logistics company that designs, manufactures, and integrates logistical systems. It provides solutions for order fulfillment and packing and storing/retrieval of unit loads.

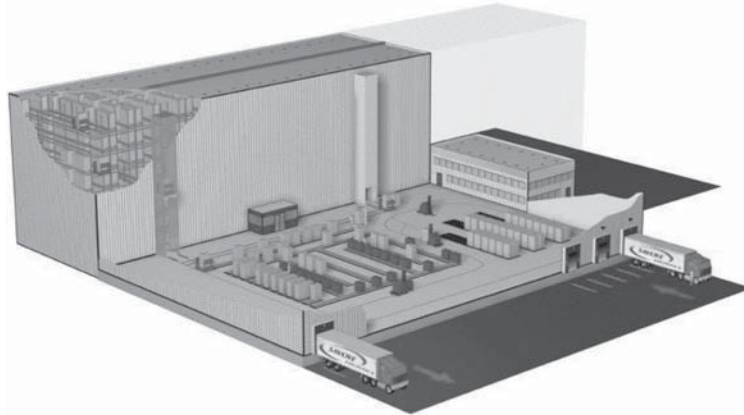


Figure 5 Three-dimensional view of a warehouse with an AVS/RS.

Savoie has various teams to assist its customers with logistic expertise to provide them the best solution corresponding to their needs. Their aim is to guarantee the performance to the customer by selecting the best equipment and a global management of the entire project.

Savoie Logistics has introduced the AVS/RS shown in Fig. 5. The system has been successfully installed in 35 companies in Europe. Today, the installed systems' capacities are around 1,000,000 pallets and 100,000 movements per day in eight countries.

There are two components of the AVS/RS, one of which is an autonomous vehicle and the other is the lift. Although the autonomous vehicle moves horizontally in the storage areas in a given tier, the lift moves the vehicle between tiers. In other words, autonomous vehicles move on rails in the aisles and interface with lifts for vertical movement of pallets between storage tiers. Here, lifts are like conveyors, but they can travel only vertically. Autonomous vehicles also transport pallets between lifts and shipping/receiving areas at the ground level. They can transform the loads from their stored areas to their respective storage addresses in the same tier because they can move within tiers. If the load movement is not on the same tier, then lifts are used for transferring the load to the related tier.

The different load movement patterns make AVS/RSs more flexible than AS/R systems, although at slightly lower efficiency. In AS/RSs, aisle-captive cranes are the main S/R devices to move unit loads simultaneously in the horizontal and vertical dimensions. Unlike storage cranes in AS/RSs, AVS/RS vehicles can access any designated storage address but must move in a sequential, rectilinear pattern.

One of Savoie Logistics' AVS/RS applications completed in 1999 was for a telecommunication company that faced rapid growth in one of its warehouses. The logistical challenge in planning was to link the technological manufacturing levels of two buildings with a production supply chain to sustain the material flow from the assembly lines to dispatch. The AVS/RS designed by Savoie was able to satisfy these constructional requirements with a supply and unloading line offset at an angle of 90°. The system has now been in operation and fulfills the short lead times and safety requirements and can achieve fill rate of 95%.

REFERENCES

1. J. M. Apple, *Plant Layout and Material Handling*, 3rd ed., Wiley, New York, 1977.
2. F. E. Meyers, *Plant Layout and Material Handling*, Regents/Prentice Hall, Englewood Cliffs, NJ, 1993.
3. S. S. Heragu, *Facilities Design*, 3rd ed., CRC Press, Clermont, FL, 2008.

4. S. Sujono and R. S. Lashkari, "A Multiobjective Model of Operation Allocation and Material Handling System Selection in FMS Design," *Int. J. Prod. Econ.*, **105**, 116–133, 2007.
5. J. Paulo, R. S. Lashkari, and S. P. Dutta, "Operation Allocation and Materials Handling System Selection in a Flexible Manufacturing System: A Sequential Modeling Approach," *Int. J. Prod. Econ.*, **40**, 7–35, 2002.
6. R. S. Lashkari, R. Boparai, and J. Paulo, "Towards an Integrated Model of Operation Allocation and Materials Handling Selection in Cellular Manufacturing System," *Int. J. Prod. Econ.*, **87**, 115–139, 2004.
7. R. U. Ayres, "Complexity, Reliability, and Design: Manufacturing Implications," *Manufact. Rev.*, **1**, 26–35, 1988.
8. J. D. C. Little, "A Proof for the Queuing Formula $L = \lambda W$," *Oper. Res.*, **9**, 383–385, 1961.

PART 2

**MANAGEMENT, FINANCE,
QUALITY, LAW,
AND RESEARCH**

CHAPTER 17

INTELLIGENT CONTROL OF MATERIAL HANDLING SYSTEMS

Kasper Hallenborg
University of Southern Denmark
Odense, Denmark

1 HISTORICAL INTRODUCTION	516	11 CASE STUDY 1: BAGGAGE HANDLING SYSTEM	531
2 FLEXIBLE MANUFACTURING	516	11.1 Performance Criteria	532
3 DISTRIBUTED SYSTEMS	517	11.2 Worst-Case Scenario	533
4 NEW CHALLENGES	519	11.3 Agent Design	534
5 AGENT TECHNOLOGY	520	11.4 Toploader	534
6 MULTIAGENT SYSTEMS	522	11.5 Agent Interactions and Ontology	538
7 AGENT TYPES	523	11.6 Internal Agent Reasoning	540
8 AGENT ARCHITECTURES	523	12 CASE STUDY 2: MATERIAL HANDLING IN AN ANODIZATION SYSTEM	543
9 AGENT COMMUNICATION	526	12.1 PACO Approach	545
10 AGENT ORGANIZATION	528	12.2 Agent Design	546
10.1 Hierarchies	529	12.3 Interactions	548
10.2 Holarchies	529	13 RESULTS	550
10.3 Coalitions	529	13.1 Active, Sleeping, and Locked Agents	550
10.4 Teams	530	13.2 Predecessor Validation	552
10.5 Societies	530	13.3 Floating	552
10.6 Federations	530	14 SUMMARY	554
10.7 Markets	530	REFERENCES	555
10.8 Matrix	531		

Manufacturers in highly developed countries around the world have, during the last decade, experienced new challenges due to globalization and changes in customer requirements. Shortening the time to market for new products and user customization are some of the factors that challenge production planning for many companies. Mass production of highly standardized products either is moving to low-wage countries or is being replaced by more sophisticated alternatives required by more demanding consumers.

Mass customization is the new concept for manufacturers, introduced by Stan Davis,¹ that challenges the traditional neoclassical economic model of customers as rational consumers who seeks to maximize their benefits and minimize their costs. The growth in communication technology, globalization, and improved economy of consumers has shifted the decision-making power from the producers and the governments to the customers.

This chapter will start with an introduction of a more decentralized approach for controlling systems for manufacturing and material handling. Different approaches for intelligent control will be discussed, and finally two cases of material handling will be presented—one large-scale complex system of baggage handling in an airport and the other a case of scheduling items through a manufacturing process.

1 HISTORICAL INTRODUCTION

Beginning with the Industrial Age, high-volume, low-variety products were the new trend among manufacturers, resulting in low-cost, high-quality products. To begin with, customers were satisfied by the new opportunities realized by mass production, even though customer requirements were not the driving force in product design. Due to low competition in markets, manufacturers were more concerned with production efficiency than customer requirements.² Likewise, dominating management theories of that time focused on rationalization, such as Taylor's scientific management.³

Improvements in automation technologies led manufactures to see the possibilities of exchanging labor-intensive tasks with specialized machines and material handling systems to rationalize production. The automotive industry was among the first to take advantage of automation: Oldsmobile Motor introduced a stationary assembly line in 1907, followed by a moving assembly line in 1913 at Ford's new factory in Highland Park, Michigan, even handling parts variety.²

For decades, mass production, automation, rationalization, and scientific management were the dominating factors in manufacturing, but that gradually changed toward the end of the twentieth century. Especially due to the growth in international competition, market demands pushed forward new challenges for manufacturing—flexibility and customization. Companies in Japan were the first to address the new conditions; they changed from mass production to lean production. Instead of focusing on having high volume and rationalization as the key drivers in developing mass production environments, lean production focuses on the whole process of production—eliminating inventory, decreasing costs, increasing flexibility, minimizing defects, and creating high product variety.

As trends in the automotive market changed, customers were no longer satisfied by standard cars but required customization.⁴ Davis¹ was the first to introduce the concept of mass customization, which tries to combine the low unit cost of mass-produced items with the flexibility required by individual customers through computerized control of production facilities.

2 FLEXIBLE MANUFACTURING

As flexibility was commonly accepted as one of the primary nonfunctional requirements for new manufacturing systems, research and development initiatives naturally concentrated on means and technologies to cope with the new demands.

The notion of a *flexible manufacturing system* (FMS) was born when Williamson in the 1960s presented his System24, a flexible machine that could operate 24 h without human intervention.⁵

Computerized control and robotics were promising tools of the framework for automation, which could increase flexibility. Obviously, not all products or systems would benefit from or require increased flexibility, but FMS was intended to close the gap between dedicated manufacturing hardware and customization, as outlined by Swamidass⁶ in Fig. 1.

FMS has the advantages of zero or low switching times and hence is superior to programmable systems. Despite that, however, FMS has had only limited success in manufacturing setups.

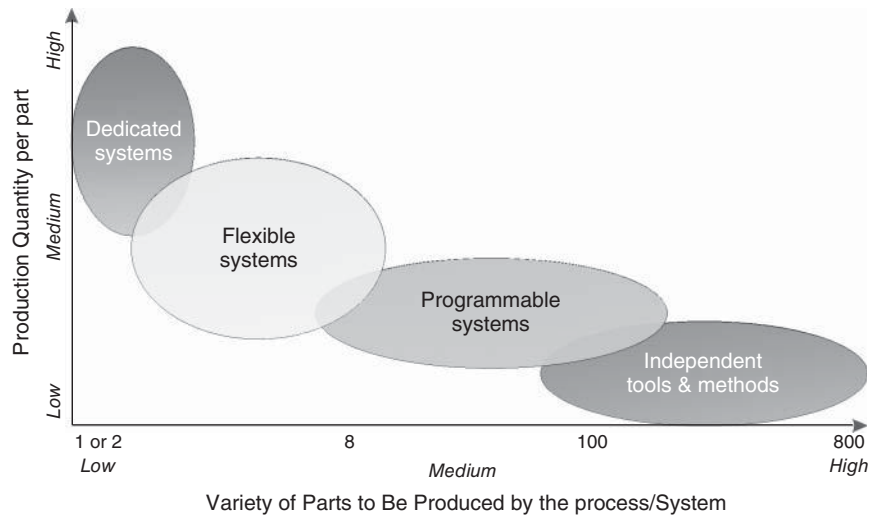


Figure 1 Manufacturing flexibility spectrum. *Source:* Adapted from Ref. 4.

Systems integration is the main issue for FMS to be successful, and flexible hardware and manufacturing entities represent only one part of the answer. The control software to handle and integrate the flexible entities in the overall process is equally important.⁴

In fact, the control software is often regarded as the critical part, as it requires high expertise from developers. The complexity of the system and time-consuming process for reconfiguration have often led to low understandability of the system, which is an important problem to manufacturers, who are not experts in manufacturing technologies.

FMSs are often composed of computer-aided or robotic assembly nodes, which are connected by some form of material handling system. Each cell can automatically handle either planned or unpredicted changes in the production flow.

The centralized control generally used in FMSs—which are based on principles and algorithms of classical control theories—would not scale very well for such large systems as identified by Sandell et al.⁷ That was the main issue leading to new approaches for manufacturing control. Bussmann (Ref. 8, p. 3) was even more specific and clear in his conclusion:

Manufacturing systems on the basis of CIM (Computer Integrated Manufacturing) are inflexible, fragile, and difficult to maintain. These deficits are mainly caused by a centralized and hierarchical control that creates a rigid communication hierarchy and an authoritarian top-down flow of commands.

3 DISTRIBUTED SYSTEMS

The experienced problems with complexity and maintenance led to new approaches in the area of manufacturing control. Parunak⁹ states that traditionally a centralizing scheduler is followed by control, which would generate optimal solutions in a static environment, but no real manufacturing system can reach this level of determinism. Even though scheduling of a shop floor environment could be optimized centrally, the system would fail in practice to generate optimal solutions due to the dynamic environment caused by disturbances such as failures, varying processing time, missing materials, or rush orders.⁴

In general, rescheduling and dissemination of new control commands are time consuming and bring the centralized model to failure. Instead, Parunak¹⁰ argued that manufacturing systems should be built from decentralized cooperative autonomous entities, which—rather than following predetermined plans—have emergent behavior spawned from agent interactions. He listed three fundamental characteristics for a new generation of systems:

1. Decentralized rather than centralized
2. Emergent rather than planned
3. Concurrent rather than sequential

The area for intelligent manufacturing systems was born, and research was conducted in different directions. One of the major approaches was a project under the intelligent manufacturing systems (IMS) program, called holonic manufacturing systems,¹¹ which settled as a new research area for manufacturing control. Holonic systems are composed of autonomous, interacting, self-determined entities called *holons*.

The notion was much earlier introduced by Koestler¹² as a truncation of the Greek word *holos*, which means “whole.” The suffix on that means “part,” similar to the notion used for electrons and protons. Thus *holons* of the manufacturing entities are parts of a whole.

The HMS project was initialized by a prestudy,¹¹ before the large-scale project was launched in the period from 1995 to 2000. The huge initiative had more than 30 partners worldwide. Not only did the project focus on applications, but also three of the seven work packages concentrated on developing generic technologies for holonic systems, such as system architecture, generic operation (planning, reconfiguration, communication, etc.), and strategies for resource management. The application-oriented foci were organized in four work packages concerning manufacturing units, fixtures for assembly, material handling (robots, feeders, sensors, etc.), and holomobiles (mobile systems for transportation, maintenance, etc.).

The project was very successful regarding generic structures of the holons aimed at low-level, real-time processing. The specification of the holons was even formally standardized by the International Electrotechnical Commission (IEC) 61499 series of standards.

The holonic parts came up short in systems requiring higher level of reasoning,¹³ thus the term *holonic agents* was introduced.¹⁴ Software agents encapsulate the holon and provide higher level decision logic and reasoning, but also more intelligent mechanisms to cooperate with other holonic agents.

Generally, agent technologies provide a software engineering approach to analyze, develop, and implement intelligent manufacturing control for distributed entities and holons. Whereas the holons were formally specified through the IEC standards, agent-based manufacturing control still lacks from having formal standards, even though various attempts have been taken—YAMS (Yet Another Manufacturing System) by Parunak¹⁵ or MASCADA,¹⁶ among others.

Research on manufacturing and material handling systems has gradually moved from a monolithic control toward a decentralized, distributed, and—most recently—agent-based approach, but only a few real systems have adopted the shop-floor models. Production 2000+ (P2000+) is an exception and is generally known as the first agent-based manufacturing system that has moved from research into real production. P2000+ was installed at Daimler to produce cylinders. The objectives of the project were to develop a robust and flexible manufacturing system through a set of flexible machines that were connected by a flexible material handling system.¹⁷ An overview of the P2000+ system layout and agent mapping is illustrated in Fig. 2.

Before going into details about agent technologies and how they can be applied to manufacturing and material handling systems, this chapter will address the issues of the introduction. The new conditions for manufacturers that push toward more intelligent control will lead to

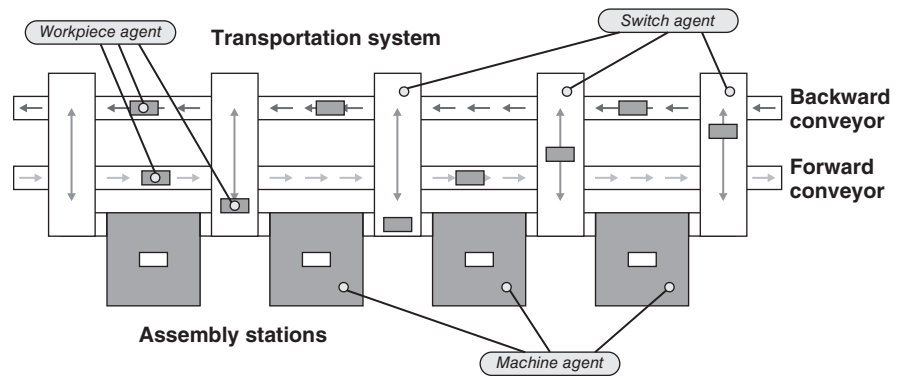


Figure 2 Illustration of the system layout and agent mapping in Production 2000+. *Source:* Adapted from Ref. 17.

new objectives for the design of such systems, and it is important to know if they fit with an agent-based approach.

4 NEW CHALLENGES

Volatile markets, globalization, emergent technologies, and increasing customer requirements are pushing new challenges to manufacturers. Shen and Norrie summarize a number of fundamental requirements that must be considered for the next generation of manufacturing systems¹⁸:

- *Enterprise Integration.* With constantly changing market and user requirements, the time to market is decreasing. Thus, a competitive manufacturing system must be integrated with related management systems, so purchasing, orders, personnel, materials, transport, and so on, are taken into account.
- *Interoperability.* The information environment for new systems can no longer be expected to be homogenous and from the same vendor. Systems may be composed of subsystems, which must cooperate and interact.
- *Open and Dynamic Structure.* New subsystems could even be added during operation, which require open and dynamic architectures that allow new components to be integrated regarding both software and hardware.
- *Cooperation.* Cooperation must be established with customers, suppliers, and other partners in order to secure the flow of materials and discover all customer requirements.
- *Agility.* Agile manufacturing is a key concept. It is the ability to quickly adapt and reconfigure the manufacturing environment to unanticipated changes.
- *Scalability.* It is important that the organization can grow or shrink at any level when required.
- *Fault Tolerance.* Environments are not static. Failures will occur once in a while; therefore the system must be able to detect and recover from such system failures.

The worldwide trend toward low-batch production with an increased variety has been ongoing for several decades. The growth and advancements in communication technologies have enabled customers to raise the individual desires or even take part in the development process. Information technology has also made it much simpler to get valuable customer feedback, which can evolve and improve the products—thereby also shorten the product life cycle.

Previously manufacturing sites were optimized by a linear production model that was suitable for mass series production, such as transfer lines, but long switching times make them inherently less suitable for flexible manufacturing, resulting in low utilization.

Both flexible and distributed systems, as presented in the previous section, can meet many of the requirements mentioned but fail on others as well. Holonic systems focus on creating flexible systems through decentralized and cooperative components, which will benefit the systems with respect to agility and scalability of fault tolerance, but the heterogeneous environment and enterprise integration are given no special attention. Distributed systems might solve the issues of interoperability, as well as open and dynamic structures, but at the local scope there is no guarantee for flexibility and efficiency.

Computer-integrated manufacturing has also been proposed as a solution to cope with the new challenges. CIM is an approach where the entire production process is controlled by a computer. It is organized in a hierarchical architecture from the strategic level of the company to the production level, but with closed-loop control so that feedback is provided back to the subsystems in order to optimize the entire process.^{19,20} However, the centralized and sequential approach to control both planning and scheduling was found insufficiently flexible and agile for the dynamic production environment and the changing production styles.²¹ Huge investments were required to install the sensory feedback at the physical machine level and implement them in the centralized monolithic control system. The complexity of the system made them rigid and inflexible, and often conditions had changed when the systems were fully operational. The organization of CIM factories is commonly hierarchical, so a single point of failure could shut down the entire system. For these reasons, the original approach of a CIM factory was never successful in real life,²² as the new requirement of flexibility, dynamic production environment, and mass customization overtook the efficiency of CIM. Instead, the responsibility in CIM systems was distributed to autonomous, intelligent, and collaborating components, which led to the holonic manufacturing approach under the IMS program already mentioned.

The shortcomings of the holonic manufacturing gradually prepared the agent-based approach as the most promising software technology for intelligent control. Whereas holonic systems have the focus on all the mechatronic components of an IMS, an agent model of the system also incorporates the planning, scheduling, and interoperability among the agents.

Multiagent-based systems (MASs) are still a relatively new paradigm in computer science, which can suit many other purposes than control of logistics and manufacturing systems. MASs facilitate an optimization of the decision process and add an extra level of robustness and stability to complex, heterogeneous systems. Agents are able to interact in dynamic, open, and unpredictable environments with many actors, here called *agents*, who cooperate to solve specific tasks or achieve design goals.

Usually, an agent-based manufacturing and material handling system is modeled with agents in all the decision points of the systems, such as assembly stations, employees, cranes, automated guided vehicles (AGVs), robotic cells, programmable logic controllers (PLCs), and so on. These agents will be able to act autonomously by observing their own local neighborhood and communicate with other agents. An agent will also act and change its actions according to the current status of its environment, so it can achieve its design goals as best as possible. This makes the system robust to local unpredictable events, as they will only be perceived by the relevant agents, who will adjust their actions, and the effect will propagate throughout the system through the interagent communication.

5 AGENT TECHNOLOGY

Agent-based, or multiagent, systems had emerged from artificial intelligence long before they were considered for control in manufacturing processes. The research area of artificial intelligence was born in the late 1950s and was focused on both understanding the human reasoning

process and developing methods and tools to build intelligent systems. In the first decade, expert systems were the primary base for research in artificial intelligent systems. The decision process of the systems was usually modeled as condition–action rules that were triggered by events from the environment or changes in an internal world model.

Pattern matching and understanding natural languages were hot topics of the time for such types of systems. It was natural to have different knowledge sources that work on different aspects of the problem, which again led to the notion of distributed artificial intelligence (DAI). Erman and Lesser²³ used different expert systems to partially process recorded data in the HEARSAY speech understanding system. They used a blackboard architecture to combine partial results to find the overall solution to the problem. The different knowledge sources each represent a different aspect or hypothesis on the problem. These are connected to the blackboard and can modify and update the current solution through the shared memory, which contains the definition of the problem.

In the following years, research led to new approaches for distributed problem solving. The contract-net protocol introduced by Smith²⁴ was a turning point in DAI. In contrast to the blackboard model, the contract-net protocol has a managing node, which through messages proposes a task to the different knowledge sources, which each bid on the task. The manager decides which (could be several) of the contracting nodes can carry out the task and eventually return the result to the manager.

Hewitt was interested in the modeling aspects of distributed problem solving and introduced the actor model.²⁵ Actors are computational entities with both a script that defines the actions and a list of other actors it can contact—the so-called acquaintances. In the model, actors are awakened when they receive messages from other actors. The actor then runs its script, will die, and is subsequently removed by the garbage collector. During execution of the script, it can both spawn new actors and send messages to its acquaintances.

Given the concepts of message passing and well-defined behavior through the action script, the actor model was a natural predecessor to the multiagent paradigm.

Multiagent systems are appropriate for studying and managing dynamical and heterogeneous systems. They are an approach to handling the increasing complexity of centralized systems by breaking them into simpler tasks, which also give a more natural modeling approach.

Starting with the simplest nonintelligent agents, there exist no commonly agreed-upon definition of an agent, but it is generally accepted that it involves some kind of autonomy, which means that the agent is allowed to choose its own action. Also, the notion of being in an environment is central to agents, as they base their actions on sensory impressions from the environment, which could be either physical or virtual environments. Thus, some sort of input function or perception unit is required for an agent. One of the most cited definitions of an agent is given by Wooldridge²⁶:

An agent is a computer that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives.

An agent can be as simple as your heating thermostats, but intelligent agents are the most interesting to be studied in both agent and multiagent systems. Figure 3 presents the classic illustration of an agent situated in its environment.

There might not be some clear distinction between intelligent or nonintelligent agents, but intelligence is usually combined with some sort of learning mechanism, and Wooldridge and Jennings add these properties to the classification of an intelligent agent²⁷:

- *Social*. Intelligent agents can interact with other agents and systems, including humans.
- *Proactive*. Agents not only respond to their environment, but they can exhibit goal-directed behavior on their own initiative.

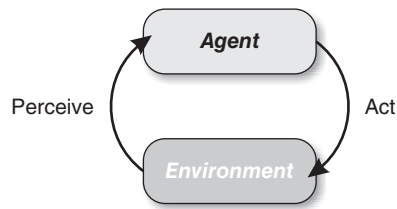


Figure 3 Classic agent illustration.

- *Reactive.* Agents are capable of perceiving changes in their environment and react upon that.

Other characteristics, such as adaptability, mobility, and rationality, just to mention a few, have been related with the agent model, but it is commonly accepted that the four concepts just described (autonomous, social, proactive, and reactive) are the main characteristics of an agent.

6 MULTIAGENT SYSTEMS

Single-agent systems are very close to traditional centralized monolithic systems, but the real strength of agents is revealed when multiple agents interact. Therefore a MAS is defined as follows: A multiagent system is a collection of interacting agents.²⁸

It is important to state that a MAS is far more than just a collection of agents, as it also encompasses the emergent behavior, which spawns from the organization of agents and their interaction under the influence of the environment in which they are situated. Da Silva and Demazeau²⁹ proposed the *Vowels* formalism to model a multiagent system:

$$\text{MAS} = \text{A}(\text{Agents}) + \text{E}(\text{Environment}) + \text{I}(\text{Interactions}) + \text{O}(\text{Organization})$$

- *Agents.* Agents are the classic entities to consider when developing a MAS system as agents are determined to be the local actors carrying out the tasks of the system. Agents are the key concept that comprises the system, which follows the characteristics of single agents.
- *Environment.* The environment is the space in which the agents exists, moves, and interacts. The space could be virtual, informational, and conceptual, but typically the environment is represented by a model of the physical space of the MAS community.
- *Interactions.* Interactions and communication are evident in MASs due to the aspects of distribution in the systems and originally by Wooldridge and Jennings²⁷ as the social ability of an agent. Interactions in the MAS community could take many forms: negotiation, collaboration, coordination, queries, or generally any kind of information exchange between agents. Interactions could be formed as abstract speech-act messages or in other models as simple natural forces that influence other agents.
- *Organization.* Similar to humans, agents can benefit from being organized, either explicitly defined in classic organizational structures, or the organization could emerge from simple interactions among the agents. The organization often serves the purpose of grouping agents with similar or related actions or behaviors. Organizations can be helpful to support agents in planning, performing actions, requesting information, or realizing global goals of the agent system.

In the research of multiagent systems, two different perspectives are dominating—either a microlevel or macrolevel perspective. For example, the system could focus on the microlevel

issues, such as the internal of the individual agents. What is the decision logic of the agent, how will the agent learn, and how can it be ensured that the agent will act autonomously? In the macrolevel perspective the multiagent research community is more concerned with the organization of agent and how the agent will interact and collaborate in an efficient way. Whereas the microlevel perspective is commonly inspired by biological systems, such as the human brain and ant colonies,³⁰ the macrolevel perspective analogies are coming from human organizations and societies.^{31,32}

7 AGENT TYPES

Agents are usually classified as being either *reactive* or *deliberate* in their behavior.²⁸ Coming out of the artificial intelligence community, a deliberate or cognitive behavior of agents was expected:

- *Cognitive or Deliberative Agents.* A cognitive agent is one that owns a knowledge base and holds a model of the current environment as it has perceived it, but it will not act only on a search in the knowledge base. It has planning capabilities, so it proactively can adjust its actions according to its goals, even though the perceived input is not covered by the knowledge base.

For the reasoning capabilities, a cognitive agent needs a representation of itself, the part of the environment in which it operates and exists, but also the agents with which it has to communicate. Thus, the internal state given by this world model will very likely influence the decision and current actions of the agent. The BDI architecture by Rao and Georgeff³³ is the most classic example of a cognitive agent model, based on desires, beliefs, and intentions, which will be described next.

For cognitive agents, their actions should not be seen as direct action of the changes they perceive from the environment, but more as a result of their reasoning on understanding the world in which they are situated.

- *Reactive Agents.* For reactive agents, there is expected to be a matching rule of action in the knowledge base for each of the inputs it perceives that will lead to actions of the agent. Actions are a direct reaction on the inputs of an agent. A reactive agent usually has no internal world model, as its actions are fully described by rules or functions of the input. The knowledge base for reactive agents could be a rule base known from expert systems, where conditional rules map to a specific output, or physical or biological inspired functions, such as physical forces—which again impact the environment or other agents. Therefore, reactive agents are less social and have only little or no direct communication with other agents. Their communication is more indirect through the environment, such as the concept of stigmergy in swarm intelligence.³⁴

8 AGENT ARCHITECTURES

The agent architecture is closely related to the agent type, as the architecture in the internal organization of the agent, which describes how it reasons and reacts to perceived input. The architecture presents a design model of the agent, where the flow of information from input to actions are explicitly defined through basic concepts of the agent, such as perception, goals, and desires.

A number of different architectures have been proposed for agents, and they are often associated with the type of agents that participate in the systems. One of the best examples of an architecture that support the reactive behavior of the agent is the subsumption architecture by Brooks.³⁵ The principle of the architecture is that an agent has a set of accomplishing behaviors arranged in a subsumption hierarchy. Each of the behavior maps a given set of input values

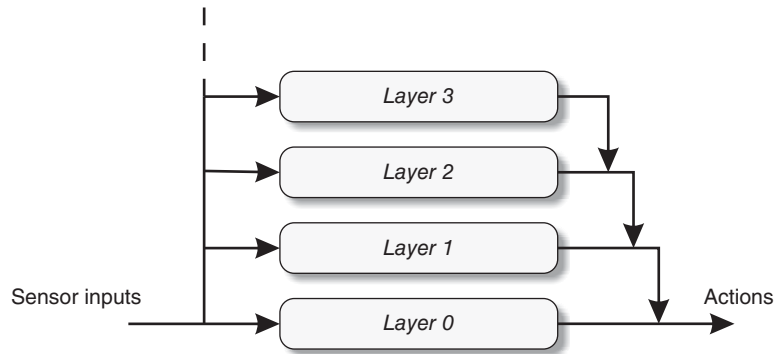


Figure 4 Brook's subsumption architecture.

directly to an output value that affect the actuators of the agent. The behaviors in the hierarchy are arranged so lower layers represent low-level behaviors and are prioritized over higher layers that represent more abstract behaviors. The subsumption architecture has been found useful for controlling robots and other AGVs, where lower layers represent the basic high prioritized tasks, such as obstacle avoidance, while the upper layers focus on the general goals of the robots, such as going from A to B or exploring an environment. The famous boid model of Craig Reynolds simulating the flocking behavior of birds and fish use a similar approach to economize with the energy consumption of the boids.³⁶ Figure 4 shows a model of the layered subsumption architecture by Brooks.

For cognitive and deliberative agents, the most well-known architecture that comprises most the concepts and ideas of highly reasoning and cognitive agents is the BDI architecture.³³ Here, B is for beliefs, D for desires, and I for intentions, which very well reflect the principles of human reasoning and other intelligent creatures. The agent has a current view of the world or the environment in which it is situated, which is modeled through its beliefs. The goals it has been designed to achieve are described by the desires of the agent. One can think of the desires as a plan library or a set of described goals that the agent wants to achieve but not necessarily is working on at the moment. So the decision making of the agent works by selecting the desires that seem most achievable under the current conditions (the beliefs). When an agent commits to pursuing a certain desire, the desire becomes an intention of the agent, and the agent persists in pursuing this goal until it no longer appears achievable. Thus, the agent will not just give up on a current plan, whenever new inputs are perceived, so that challenge of using the BDI architecture is to balance the proactive and goal-directed behavior against the influence from new inputs, which is a more reactive behavior. A model of the BDI architecture is presented in Fig. 5.

Most other architectures for cognitive agents either are an extension of the BDI architecture or use a somewhat similar approach with a formalized reasoning model between a set of described goals and the current world model.

Hybrid architectures, which try to combine the best of both worlds, also exist, and the InteRRaP architecture (Fig. 6) by Müller is a classical example of that.³⁷ The InteRRaP architecture has three layers:

1. A bottom layer has all the reactive behaviors and situation-to-action rules. This layer is also the interface to the world.
2. The plan layer is on top of the behavior layer. It handles all the goal-directed and proactive planning of the agent.
3. A cooperation layer describes the collaboration and interactions with other agents.

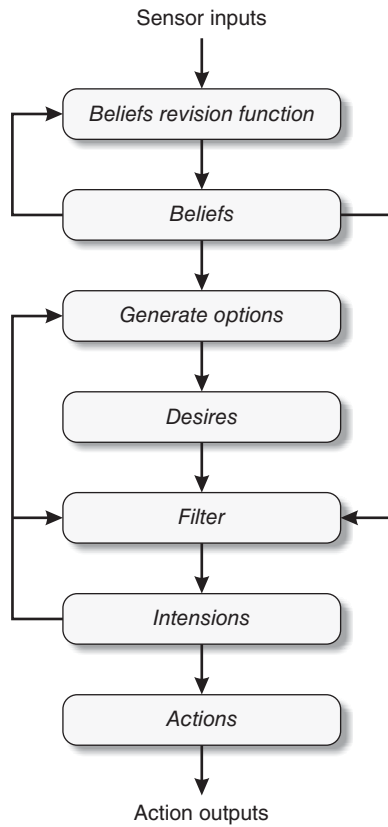


Figure 5 BDI architecture.

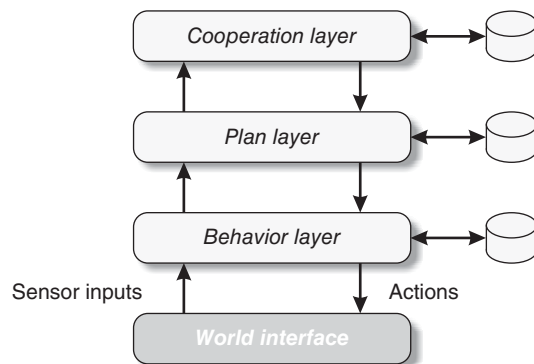


Figure 6 InteRRaP architecture.

A common problem with hybrid architectures such as InteRRaP is that there is no clear semantics or methodology for programming such architectures,³⁸ so it can be hard for the developer to design a coherent agent behavior.

9 AGENT COMMUNICATION

Interagent communication is a key requirement for an agent to fulfill its social characteristics, so a long list of interaction schemes and communication protocols have been proposed and implemented in the agent community. However, with respect to manufacturing and material handling systems, particular coordination, negotiation, and hybrid mechanisms are dominating.²⁸

Apparently, coordination is a form of communication—not only for agents—that will be most noticeable if it does not work properly. Perfectly aligned conveyors, input, and output facilities of material handling systems require a high degree of coordination between the control elements, which are obtained through an often long and tedious alignment process at installation time. For flexible manufacturing systems, the conditions for coordination are constantly subject to changes. Therefore, coordinating agent activities is of highest priority in intelligent manufacturing and material handling systems. With a system composed of flexible cells and connected through, for example, conveyors, coordination can be defined as the process of managing dependencies between activities.³⁹ Coordination is not a trivial thing to achieve in agent-based control systems. Jennings has emphasized three common characteristic of agent systems that lead to dependencies and complicate coordination⁴⁰:

- *Actions of agents might interfere.* Two agents might fight for the same resource in order to complete their tasks.
- *Global constraints might have to be satisfied.* It could be suitable to distribute the load on the entire system, but there will still be an overall deadline of an order that an item must satisfy.
- *Individual agents cannot satisfy their own or system goals by itself.* The core idea of a flexible production cell is that an item has to process several work stations for it to be completed.

Malone and Crawston³⁹ further simplified the interdependencies relevant for distributed agents into three types of dependencies, as shown in Fig. 7. A *flow dependency* arises if one task in the process produces or generates a resource that is required by another task. This is the most common dependency in material handling systems (i.e., one conveyor transfers an item to the next conveyor). *Sharing dependencies* occur where several tasks want to access the same resource. It could be two conveyors diverting from a single conveyor line, where the diverter is the sharing resource that is part of both tasks. A *fit dependency* exists when two or more tasks collectively produce a single resource. The obvious example is, of course, the composition of a complex item, but in a material handling system, fit dependencies also exist when two conveyor lines merge into a single line.

Modeling of dependencies and task specifications has been formalized by Decker in the Task Analysis, Environmental Modeling, and Simulation (TAEMS) model.⁴¹ TAEMS

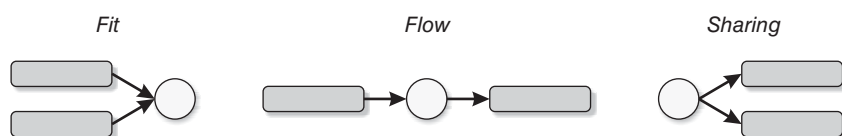


Figure 7 Dependency types between task and resources. *Source:* Adapted from Ref. 39.

is a framework to model complex computational task environments, either for formalized agent systems or experimental examples. Agents are represented in TAEMS as the executing, communicating, and information-gathering entities, and a tree diagram can be drawn to visualize the dependencies and associations among agents.

Decker also proposed generalized partial global planning (GPGP) as a set of generic coordination mechanisms that can bring a decomposed distributed task model of a complex and dynamic environment (e.g., modeled by TAEMS) into a global plan for the collaborating agents. GPGP has been applied to a number of problems related to scheduling, planning, and resource optimization.^{42–44} The GPGP approach has its strengths in reactive planning for agents which are situated in a dynamic environment. The system should quickly be able to respond to such changes. Each agent optimizes its local plan and synchronizes it with the global plan using a set of standardized coordination mechanisms.

A number of other research papers deal with other approaches for coordination in a complex task environment for agent-based manufacturing.^{45,46} Primarily, the research has focused on scheduling flexibility and resource planning in the productions environment under the constraints the processing steps of the different orders have to go through.

Negotiation is the other category of communication principles that are usually applied in agent-based systems for manufacturing and material handling. Negotiation is a well-known concept in sociology for human interaction. Negotiation has motivated many approaches and interaction mechanisms for agents as well. In general, it is about a mutual agreement between two or more actors on some sort of conflicting intensions. Such conflicting interests are what really make humans unique and intelligent, compared to rationally designed systems. Normally, such conflicts can easily be resolved and we learn from it. Pruitt⁴⁷ has provided a clear and also general definition of negotiation:

Negotiation is a process by which a joint decision is made by two or more parties. The parties first verbalize contradictory demands and then move towards agreement by a process of concession or search for new alternatives.

Naturally, with this relation to social behavior in human societies, many of the negotiation mechanisms have been formalized, adopted, and extended for use in multiagent societies.²⁸

The contract-net protocol mentioned in the introduction was one of the first examples of such a mechanism,²⁴ where the agent (the manager) that wants a task to be executed proposes it to several other agents capable of performing the task (the contractors). The agreement will be a joint decision between the manager and the contractor with the lowest offer to carry out the task.

Many of the negation principles are inspired or based on microeconomic principles that provide a formal specification and rational approach to reach a joint decision. Different auction principles are common in negation in agent systems, usually modeled as one agent having a task for one or more agents that calculate a bid to carry out the task. Common auction methods are single-side, two-side, continuous, and English auctions, where the real design challenge is to find the right interests and goals of an agent to give a fair bid.

Communication principles—such as queries, requests, and publisher–subscriber relationships—are also commonly seen in agent systems, and MAS research has proposed a number of formalized approaches to specify the content of interactions. In general, all agent communication is regarded as message-based interaction, and message content is often presented in an abstract form according to speech–act theories.

For several years, Knowledge Query and Manipulation Language (KQML) was the preferred communication language supported by many agent platforms. KQML was initially specified by the DARPA Knowledge Sharing Effort, led by Finin et al.,⁴⁸ and KQML was launched as an interface to knowledge-based systems.

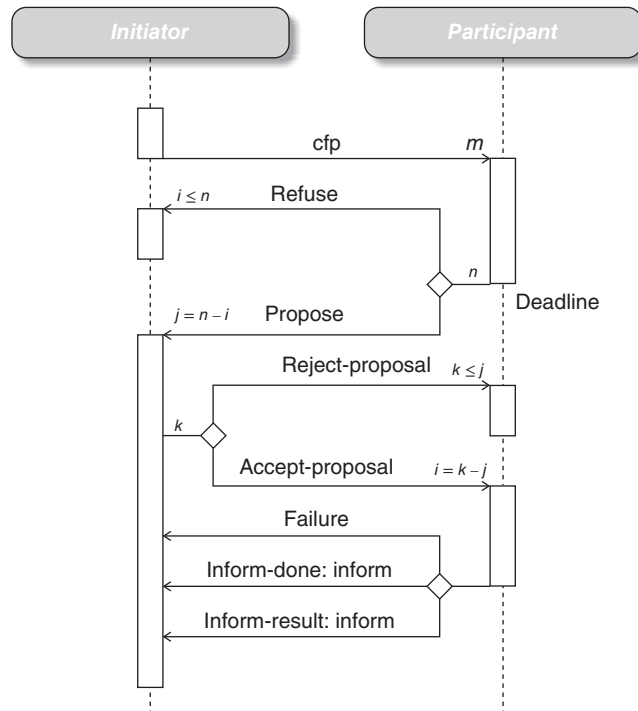


Figure 8 FIPA contract-net protocol.

Later, when the standardization organization for intelligent agents, FIPA (Foundation for Intelligent Physical Agents), announced a number of formal specifications for the agent community, ranging from specifications for interaction mechanisms to agent platform architectures, it also specified a new communication language based on the speech-act theory of Searle.⁴⁹ The language is commonly known as FIPA-ACL, which—beside the usual message information, such as the receiver and sender—also allows the message to contain information about the ontology used for encoding the content of the message, a time-out indicating the period the sender will wait for a reply, and a performative for the message, which indicates which communication act the message follows. The content field of the message is specified in a semantic language, FIPA-SL. An example of a FIPA specification for the contract-net protocol is given in Fig. 8.

10 AGENT ORGANIZATION

Whereas the architecture is an internal organization of the components and structure of the agent, organization in MAS will normally refer to a model of the structure and associations among the agents. Horling and Lesser⁵⁰ studied organizational models for multiagent systems in an extensive survey, where they state:

The organization of a multiagent system is the collection of roles, relationships, and authority structures which govern its behavior.

They also conclude that no single organizational model will suit all multiagent systems, so they present a list of organizational styles, which are briefly described in the following and illustrated in Fig. 9.

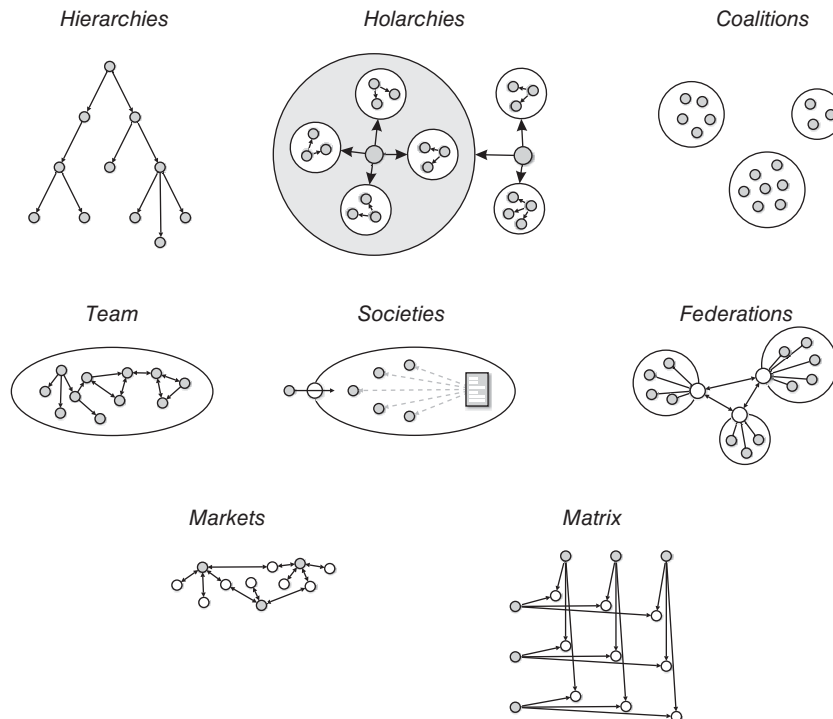


Figure 9 Organizational styles. *Source:* Adapted from Ref. 50.

10.1 Hierarchies

In a hierarchical model, the agents are arranged in a treelike structure, where direction of action comes to the leaf agents from higher level agents that have a broader view of the system. Leaf agents collect and provide information for higher level agents, and horizontal communication is usually not allowed. The strength of the hierarchical style is that parallelism is easy to achieve, and the communication flow is rather limited. The centralized characteristics of the decisional nodes also make it vulnerable for single point of failures.

10.2 Holarchies

Holarchies are based on the structural concept of encapsulation known from the object-oriented paradigm, so almost any entity can be regarded as part of something bigger that can act as a whole. For example, a wheel is a whole by itself but can also just be regarded as part of a car, and when we deal with the car, it will indirectly affect the wheel. Holarchies can be appropriate to model structural decompositions with autonomy. For manufacturing and material handling it could be different areas of a system, which have different responsibility for the process. Obviously, it decreases the predictable behavior of a holon from the outside.

10.3 Coalitions

Coalitions are another structural form known from military and business organizations, where a number of entities join forces on at least a temporary basis. Usually, there will be no internal

structure of a coalition, just a flat hierarchy of entities that coordinate their activities to pursue a commonly agreed-upon goal. Coalitions dissolve automatically when they are no longer needed or a critical number of entities leave the coalition. For material handling coalitions could be used during peak times and/or in case of breakdowns, where parts of the system join forces to cope with the current situation.

10.4 Teams

Teams are also a group of entities that cooperate and coordinate their activities to achieve a common goal, but on a more long-term or even permanent basis. Among the team members there will be some clear representation of the common goals and mutual beliefs that are fundamental for their joint work. The internal organization of a team would typically be a flat hierarchy. On the one hand, a team has the capability to handle larger jobs than a single entity, but on the other hand, communication will increase due to the internal communication inside the team. Teams would also be a typical construct in a material handling system, where parts of the system could work together in bringing an item from A to B.

10.5 Societies

In a society the behavior of agents is governed through a set of social laws, norms, and conventions. Societies are a long-term construct which groups a number of agents that can have quite different goals and heterogeneous capabilities, and the communication will usually also be more diverse and complex. The agents might require extra social skills and are typically more deliberate than simple coordinating agents. At least from the outside, it is a nice feature that the set of rules and norms are more formalized, so the behavior of agents is more predictable. Societies are not a style commonly used in the control of material handling. It would be more suitable to secure flexible interoperability on the enterprise level.

10.6 Federations

A group of agents could form a federation by selecting one group member to represent the group. All communication will go through the delegated agent that will act as a gateway to the outside. It is the responsibility of the delegate to represent and know the individual interests and capabilities of all members and incorporate them into communication with outside agents. The delegate is also commonly referred to as a broker, facilitator, or mediator. Again, federation is a common style to handle subsystem interoperability by adding an extra agent with the delegate role. The natural disadvantage is, of course, that the delegate will become a candidate for bottlenecks and a single point of failure.

10.7 Markets

Markets are based on the producer versus consumer or buyer versus seller agents principle, where one group of agents (buyers or consumers) places bids on shared resources, tasks, or services (the producers or sellers) and the best incoming offer will be chosen. In a market, agents are designed to be competitive, with the potential risk of malicious behavior among the agents, but fairness could also be increased by repeated bidding. The market style is very common for manufacturing systems, as it is rather easy to set up a price-calculating function that contains all the relevant factors to prioritize a task, such a deadline, processing time, competences of staff, and similarity with previous items.

10.8 Matrix

A matrix organization is also a common construct in human organizations, where the *worker* agents might have several relationships to different groups or managers. The style is appropriate for project organizations, where the workers belong to different functional groups but at least part time participate in projects led by other managers. The disadvantage is, of course, the potential risk of conflicts, where a worker agent has more managers, but the advantage is that capabilities of the agents can be shared and benefit several places.

11 CASE STUDY 1: BAGGAGE HANDLING SYSTEM

It might be clear from the previous sections that multiagent systems span different dimensions on how to classify the system, and in most cases the systems are hybrids, where some part of the system might contain highly cognitive agents that communicate a lot as part of their reasoning process, while other parts of the system use simple reactive agents that might solve more trivial tasks.

The two real-life examples that will be presented in this chapter well represent (for real applications) extremes in the space of these dimensions. The baggage handling system (BHS), which will be described first, is a complex system of many collaborating and negotiating agents with a cognitive behavior, where the actions of the individual agent are highly dependent on the results of communication with other agents in the systems.

Handling of baggage in airports is shadowed by matters of complexity and uncertainty from the perspective of most passengers, similar to all other issues related to air traffic. Many passengers, frequent or not, feel the moment of uncertainty when watching their bags disappear at check-in counters. Will they ever see their bags again?

Only few imagine the complex system that handles the bags in major airport hubs. Small airports or charter destinations do not fall into this category, but airports with many connecting flights experience this huge sorting and distribution problem. Baggage from check-in is usually not the biggest problem, as the sorting can, to some extent, be handled by distributing flights correctly at the check-in counters. However, bags from arriving planes that have not met their final destination will arrive totally unsorted. So the core task of a baggage handling system (BHS) is to bring each piece of baggage from the input facility to its departure gate. The identity, and hence the destination, of the bags is unknown by the system until scanned at the input facility. This makes the routing principle more attractive than scheduling and offline planning.

A BHS is a huge mechanical system, usually composed of conveyorlike modules capable of transferring totes (plastic barrels) carrying one bag each. The investigated BHS has more than 5000 modules each, with a length of 2–9 m running at speeds of 2–7 m/s. The conveyor lanes of the modules that make up the BHS in the airport of Munich, Germany, are up to 40 km in total length, and the system can handle 25,000 bags per hour, so the airport can serve its more than 25 million passengers yearly, and the BHS in Munich covers an area of up to 51,000 m². Thus, the BHS of Munich is slightly larger than the BHS presented in this case as it has 13,000 modules and more than 80 different types of modules are used, but in setup and control, they are very alike. Later, the different types of modules will be explained when describing how agents have been mapped to the BHS. Figure 10 shows a snapshot into a BHS, where a tote containing a bag runs on the conveyors in the foreground.

A BHS often covers an area similar to the basements of the terminals in an airport, and tunnels with pathways connect the terminals. The system is rather vulnerable around the tunnels, because typically there are no alternative routes and the tunnels contain only one or two FIFO-based lanes, which could be several kilometers long. Therefore, the topology of the



Figure 10 Snapshot into a BHS with a moving tote in the foreground.

BHS could look like connected clusters of smaller networks, but within a terminal, the network of conveyors is far from being homogeneous. Special areas, to some degree, serve special purposes.

Besides the physical characteristics of the BHS a numbers of external factors influence the performance:

- Arriving baggage is not sorted but arrives mixed from different flights and with different destinations, as baggage for baggage claim is usually separated and handled by other systems.
- Identity and destination of bags are unknown to the system until the bag is scanned at the input facilities; thus, preplanning and traditional scheduling are not options.
- Obviously, the airport would try to distribute the load of not only baggage, but all air-traffic-related issues over the entire airport. However, changes in flight schedules happen all the time, due to both weather conditions and delayed flights.
- Most airports have a number of peak times during the day, and flight schedules may also differ on a weekly basis or the season of the year. Peak times may influence the strategy on routing empty totes back to the inputs, as they share the pathways of the full totes.

11.1 Performance Criteria

A top priority of a BHS is that no bags are delayed, which can postpone flights and result in airports being charged by airline companies. Therefore, the BHS must comply with the maximum allowed transfer times, in this case between 8 and 11 min, depending on the number of terminals to cross. Keeping the transfer time low is a competitive factor among airports, as airline companies want to offer their customers short connections.



Figure 11 States of a bag in the BHS.

Besides ensuring that bags reach their destination in time, the capacity of the BHS should also be maximized, and the control system should try to distribute the load and utilize the entire system if it should be capable of handling peak times. Robustness and reliability are also of top priority, as breakdowns and deadlock situations inevitably lead to delayed baggage and, in the worst case, stop the airport for several hours.

To fully understand the importance of delayed bags, the concept of *rush bags* must be introduced. Dischargers are temporarily allocated to flights, which define a window where bags can be dropped for a given flight. Normally, the allocation starts 3 h before departure time and closes 20 min before departure. Bags arriving later than 20 min before departure will miss the characteristic small wagon trains of bags seen in the airport area. Thus, the system must detect if the bag will be late and redirect it to a special discharger, where these rush bags are handled individually and transported directly to the plane by airport officers. Obviously, this number should be minimized due to the high cost of manual handling.

Bags entering the system more than 3 h before departure are not allowed to move around in the system waiting for a discharger to be allocated. They must be sent to temporary storage—*early baggage storage* (EBS). Figure 11 illustrates the system life cycle of a bag with the mentioned phases.

Given those criteria, the traditional approach for controlling a BHS uses a rather simplified policy of routing totes along static shortest paths. The *static shortest path* is the shortest path of an empty system, but during operation, minor queues are unavoidable, and they lengthen the static shortest routes. In the traditional control, all totes are sent along the static shortest routes, irrespective of the time to their departure, in order to keep the control simple and reliable. A more optimal solution would be to group urgent baggage and clear the route by detouring bags with a distant departure time along less loaded areas.

On top of the basic approach, the control software is fine tuned against a number of case studies to avoid deadlock situations, but basically it limits the number of active totes in different areas of the system. The fine-tuning process is time consuming and costly for developers; hence, a more general and less system specific solution is one of the ambitions with an agent-based solution.

Naturally, the control of the BHS should try to maximize throughput and capacity of the BHS, which is indirectly linked to the issues of rush bags. Besides that, a number of secondary performance parameters apply as well, such as minimizing energy consumption of the motor and lifetime of the equipment—for example, by minimizing the number of start and stops of the elements and avoid quick accelerations.

11.2 Worst-Case Scenario

Apparently from the descriptions given here, there should be opportunities for improvement of the control logic in the BHS, and one might ask why it has not been tried before⁵¹:

Still listed as one of the history's top ten worst software scandals are the BHS of Denver airport in Colorado, US. The Denver International Airport was scheduled to open in October 1993, but caused by a non-working BHS the opening of the airport was delayed in 16 months costing \$1 million every day. When it finally opened in 1995 it only worked on outbound flights in one of the three terminals, and a backup-system and labour-intensive system was used in the other terminals.

The original plan for the BHS developed and built by BAE was also extremely challenging, even compared to many BHSs built today. Instead of moving totes on conveyors, the BHS in Denver is based on more than 4000 autonomous DCVs (destination coded vehicles) running at impressive speeds of up to 32 kph on the 30-km-long rail system. It was a kind of agent-based system with many computers coordinating the tasks, but the first serious trouble was caused by the overloaded 10M-bit ethernet. Also, the optimistic plan of loading and unloading DCVs while running caused DCVs to collide, throwing baggage of the DCVs and sometimes damaging baggage. The original plan even called for transferring baggage from one running DCV into another, whereas many systems today still stop a tote or DCV before unloading, even at stationary discharging points.

11.3 Agent Design

The *elements* are the building blocks of the BHS and from an intuitive point of view are the potential candidates for agents in the system because all actions of the system are performed by the elements. The elements are the module the BHS consists of, mostly conveyor module, which varies 2–9 m in length (some straight, some curved), but they could also be a module that can tilt or that split. But they are part of the lanes where the baggage moves around in these barrels, called totes.

The applied approach concentrates on the reasoning part of agents and their interaction from a macrolevel perspective. An alternative approach would be to consider the totes as “consumer” agents and the BHS as a collection of “producer” agents, where the BHS can solve the tasks that the totes want to have performed—bringing the tote to the destination. In principle, a tote could then negotiate its way through the system, and if the timing was urgent, the bags would be willing to pay a higher price than nonurgent bags.

Such an approach often leads to other complications, such as communication overhead and complex agent management.⁴ Because the BHS generally consists of pathways of first-in-first-out (FIFO) queuing conveyors with little and often no possibilities of overtaking each other, it is more appropriate to design the agents around the flow of the BHS, which makes the elements the potential candidates for agents. The element agents should then coordinate their activities to optimize system performance and should therefore be considered as collaborative agents, rather than competitive agents.

11.4 Toploader

The input facilities of the BHS are called *toploaders*, as they drop bags into the totes from a conventional conveyor belt (see Fig. 12). Before the bag is *inducted* into the tote, it passes a scanner that reads the ID tag and destination, so the control system has exact tracking of the bag at all time.

Identity and destination of the bag are unknown until the bag passes the scanner at the toploader. The scanning initializes routing of the tote, but the short time leaves no option for global optimized planning of all current totes with replanning for every new arrival.

Basically, the task of the toploader could be decomposed into several steps. Scanning the bag, which happens automatically, has no direct impact on the control. The toploader initiates the journey of the tote on the BHS. A destination (discharger) must be set for the tote. In order to increase capacity, several dischargers are often allocated to the same flight destination.* Therefore, the toploader agents initiate a negotiation with the possible dischargers to find the best-suited discharger. The evaluation of the proposals from the dischargers is not trivially

* Due to the stopping of totes while unloading, the discharger has a lower line capacity than straight elements.

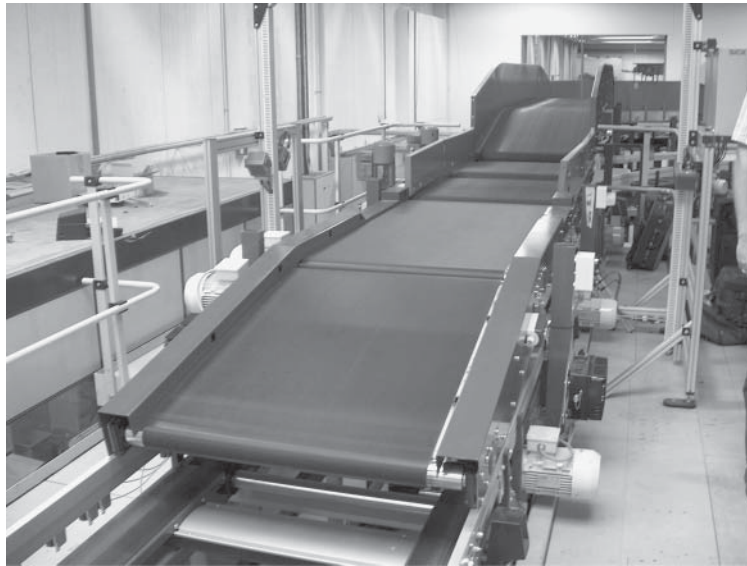


Figure 12 A toploader, where bags arrive on a traditional conveyor belt.

chosen as the lowest offer but is weighted with the current route length to the dischargers, which the toploader requests from a route agent—a mediator agent with a global focus on the dynamic route lengths of the BHS.

The toploader can take two different approaches for routing the tote:

1. *Routing by Static Shortest Path.* After the toploader has decided on the discharger, it could instruct all diverting elements along the route to direct that specific tote along the shortest path. Then the agent system would, in principle, work as the traditional control system by sending all totes along predefined static shortest routes.*
2. *Routing on the Way.* Instead of planning the entire route through the BHS, the toploader could just send the tote to the next decision point along the shortest route. This is a more dynamical and flexible approach, as the tote can be rerouted at a decision point if the route conditions have changed—perhaps another route has become the dynamical shortest one or the preferred discharging point has changed.

More formally, the principal tasks of the toploader can be illustrated as the diagram in Fig. 13, but it hides the advanced decision logic between the state changes and message interactions:

- *Straight Elements.* Most of the elements of a BHS are naturally straight or curved elements (conveyor lanes) that connect the nodes of the routing graph. Straight or curved elements are not considered as agents in our current design, because mechanically they will always forward a tote to the next element if it free; thus, there are no decisions to be made. In principle, the speed of each element could be adjusted to give a more smooth

* In the researched BHS the decision between the alternative dischargers would also be predefined in the conventional control. The BHS is built in layers to minimize cost and maximize space utilization, and alternative dischargers are always split on different layers. The control system would try to avoid switching layers.

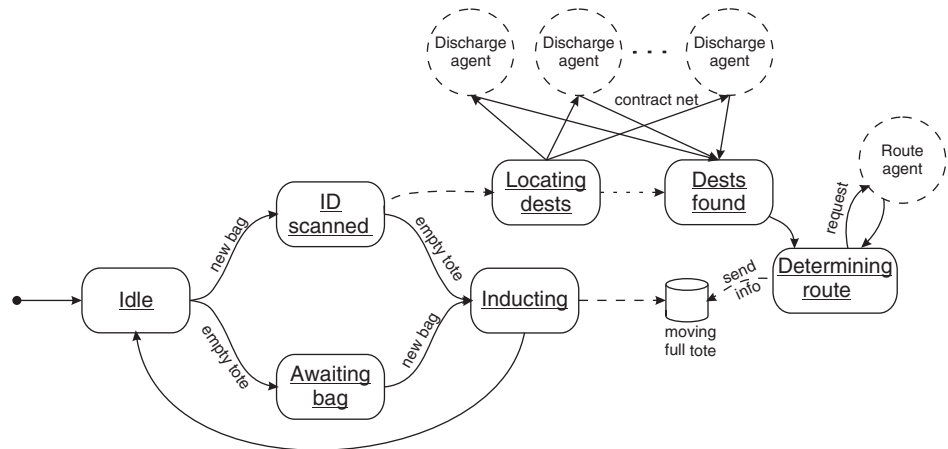


Figure 13 Principal tasks of a topper.



Figure 14 Diverter element with an empty oversize tote.

flow and avoid queuing, so one could argue that these decisions should be taken by the element itself. In the current setup, it would generate an enormous communication overhead, because each element should be notified individually and the agents should be very responsive to change the speed in order to gain anything from speed adjustments.

- *Diversers*. Diverted elements (Fig. 14) become the first natural decision points on a route. A diverter splits a conveyor lane into two, either a left or right turn and straight ahead. Lifts and so-called cross-transfers are special editions of diversers. The cross-transfer allows the tote to be forwarded in all four directions.

With respect to the strategies described here, the diverter either could just forward the tote in the direction determined by the topper or should reconsider alternative routes by restarting

the negotiation process with dischargers and requesting updated information on dynamic route lengths. A diverter should be concerned about the relevancy of reconsidering the route for a tote, because in many cases there is only one possible direction at a given diverter for a given tote.

So the decision logic, rather, is identical to the dynamic routing principle of toploaders, but diverters should fine-tune their decisions according to the local environment in which they are situated. In other words, a strong influence on the decision logic of the diverter is based on its position in the routing graph.

Similar to the toploader, the principal tasks of the diverter can be illustrated by a state diagram, shown in Fig. 15.

- *Mergers.* Mergers are the opposite of diverters, as they merge two lanes. Traditionally, mergers are not controlled, as there are no alternatives to continuing on the single lane ahead, and the merger simply alters between taking one tote from either input lane if both are occupied.

Obviously, more intelligent decisions could be considered than just switching between the input lanes, which is the argument for applying agents to the merger elements. The ratio between merging totes from the input lanes should be determined by the aggregated data of the totes in either of the two lanes (e.g., if the number of urgent totes waiting to be merged is higher in one lane, then that lane should be given higher priority). Also waiting totes in one lane could have greater impact on the overall system performance if a queue of totes in one lane is more likely to block other routes behind that point.

- *Dischargers.* Dischargers (Fig. 16) unload bags from the totes. When bags are discharged, they fall onto carrousel similar to those at baggage claim and are drawn to the plane in small wagon trains.

Besides being involved in the negotiation process described for the toploaders, the task of the discharger could seem rather simple—just tilt the tote—but a discharger also has to take care of the empty totes. Some BHSs have a separate conveyor system for the empty totes, but many systems, including the researched BHS, use the same lanes for routing the empty totes back to the tote stackers at the toploaders.

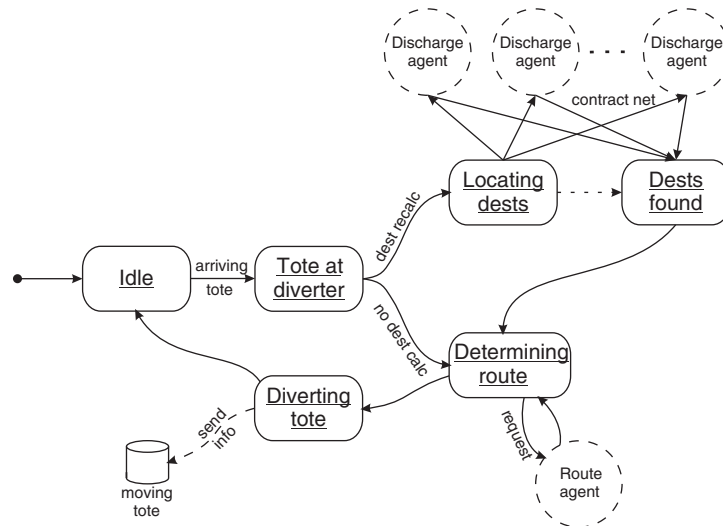


Figure 15 Principal tasks of a diverter.

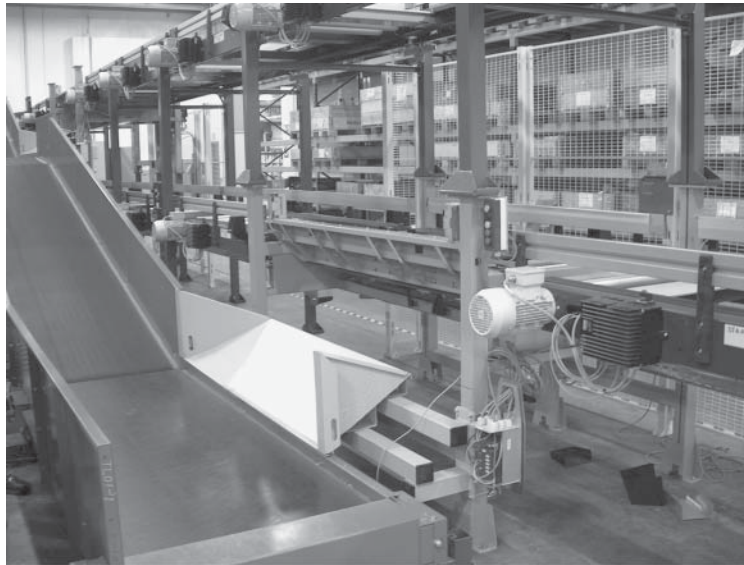


Figure 16 Discharger element that can tilt the tote, so bags slide onto the conveyor belt.

The task of routing empty totes is similar to routing full totes at the toppers but is actually much more complex due to a number of considerations that must be taken into account:

- The number of destinations (tote stackers) is larger than alternative dischargers for full totes. The number of tote stackers often matches the number of toppers, which is 12, in our case.
- Especially in the input area, empty totes are mixed with full totes, and the area could easily get overloaded and blocked.
- During peak times, some empty totes should be sent to temporary storage in the EBS, which is far from the toppers, and then released when the load on the system is lower.
- If a stacker runs empty, no totes will be available at the toppers for new bags.
- The distance to the stackers is more appropriate to return the empty tote to a stacker nearby than sending it half way through the system.

All these factors could be considered in the decision logic of the agent (e.g., by using some fuzzy set logic). The principal tasks of the discharger are illustrated by the state diagram in Fig. 17.

- *EBS Elements.* Early baggage storage elements (Fig. 18), or EBS for short, are temporary storage elements for totes with bags for which a discharger has not been allocated yet, as already described when defining the concept of rush bags.

EBS is a complete research area in itself regarding optimization of EBS elements, as totes are stored in lanes, which are released into the system again. Planning and coordinating the totes in different lanes is not a simple task but was not given further attention in the project.

11.5 Agent Interactions and Ontology

The agent interactions are based on the element's responsibility and participation in the function of the BHS, as described in the previous section. To give an example of a delegate or mediator used in the system for a federation among the agents, RouteAgent is described.

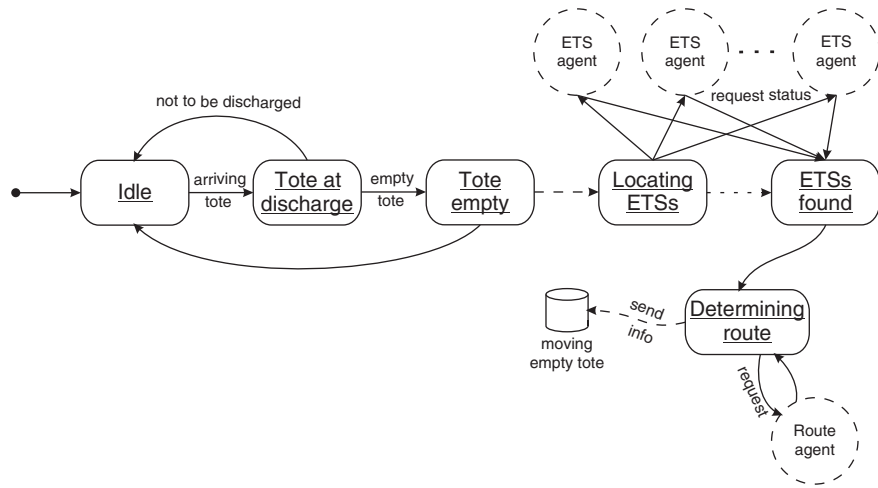


Figure 17 Principal tasks of a discharger.



Figure 18 EBS elements, here storing a line of empty totes.

There is a balance between giving agents detailed information about the environment and maintaining an internal world model or letting agents query the environment about information when required.

In theory, the interagent negotiations could be used to generate all information to route totes around in the system, but that would generate too much overhead and complicate the simple routing principles. Instead, agents can be assisted by a mediator agent that collects aggregated information for the entire system. In the initialization process, RouteAgent generates all possible routes in the system by building up a graph for the BHS with nodes corresponding to the

element agents. During the operation, it constantly monitors traffic on all edges of the graph and updated the weights in the graph, so dynamic shortest paths can be calculated using classic Dijkstra for dynamic shortest path calculations.⁵²

Following the FIPA query-ref communication act, element agents can request routes to a given destination packed in a referential expression of the query message. The referential expression is composed using an ontology that has been defined for the BHS domain, which extends and follows the structure of the FIPA-SL. RouteAgent understand two concepts of the ontology, *RouteBetween* and *LineBetween*:

1. *RouteBetween* is the concept used when agents are interested in full or parts of a route, but only with a granularity of finding other element agents along the path—only information on nodes of the graph are returned.
2. *LineBetween* is the fine-grained concept providing all details about a conveyor line of connected elements in the BHS—information about edges between two given connected nodes.

To give an example of the generality embedded in ontology-based messages, a query to RouteAgent could contain the following abstract referential expressions:

```
(iota
:Variable (Variable :Name x :ValueType set)
:Proposition (routeBetween
:origin (element :elementID DFB01.TLA001)
:destination (element :elementID DLA02.DIA023)
:viaPoints (Variable :Name x :ValueType set)
:numNodes 0
)
)
```

It is abstract because it contains the variable x that must be replaced by the responder in a response to the query. In this case, the responder is a set of points (identities of element agents between the given origin and destination). The predicate *iota* is just one of three from the FIPA-SL specification, which means exactly one object fulfills the expression, whereas the other predicates, *any* and *all*, would return any or all routes between the origin and destination, respectively.

11.6 Internal Agent Reasoning

This section will present internal agent reasoning principles to optimize the flow in the BHS in different ways to meet some of the performance parameters. Deep reasoning and long-term goals are not currently pursued in the strategies due to the flow speed and high number of totes in the system. Instead, the intentions behind the strategies are to optimize the situation for more than a single tote or forthcoming actions.

Even though the agent design does not strictly follow the BDI architecture, the behavior of the agents follows the same principle, with agents constantly monitoring the environment, and it will change its actions according to the goal it is designed to achieve based on the current state of the environment.

Three different deliberate behaviors of agent will be described which are part of both necessary routing and optimizing strategies for the BHS. The deliberate behaviors are included to exemplify how agents can have very diverse internal reasoning, which would be very hard to combine in a central solution. Intuitively, they are much easier to understand and implement when taking the perspective a single agent, its environment, and the agents with which it has to collaborate.

Returning Empty Totes

As already explained, the task of dischargers is more complicated than just emptying the tote. The tote continues on the conveyors and should be routed back to tote stackers located at the input facilities. The most important factor that influences the decision of the destination for the empty tote is the full status of the tote stackers. However, the distance to the tote stackers should also be considered. There is no reason to send it to the other end of the system if a stacker is nearby unless the other is empty.

Each stacker monitors its full status as a simple ratio between the current and maximum number of totes in the stacker. By a standard hedge⁵³ the ratio is converted into a priority s_i for requesting extra totes:

$$s_i = \begin{cases} 2r_i^2 & 0 \leq r_i < 1/2 \\ 1 - 2(1 - r_i)^2 & 1/2 \leq r_i \leq 1 \end{cases}$$

where r_i is the full-ratio for the i th stacker. A plot of the function is shown in Fig. 19.

The priority is used to scale the dynamic route length to each tote stacker, so a nearly empty stacker will have a very short route length or value in the decision, whereas a full stacker will have its full route length:

$$v_i = d_i \times s_i$$

where d_i is the dynamic distance (requested from RouteAgent) to the stacker from the decision point.

Overtaking Urgent Bags

Consider a typical layout of a discharging area in Fig. 20. The bottom lane is a fast-forward transport line, the middle a slower lane with the dischargers, and the upper lane the return path.

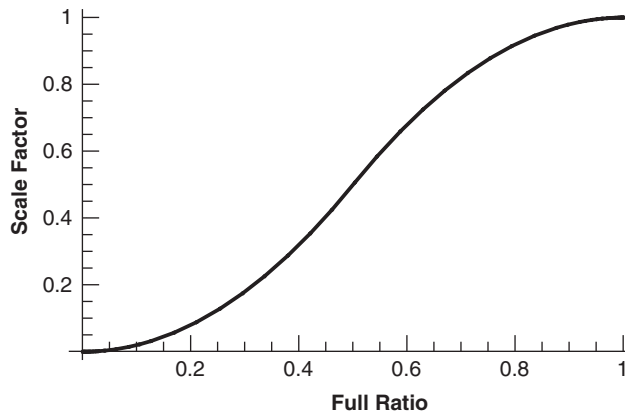


Figure 19 Plot of ETS priority function.

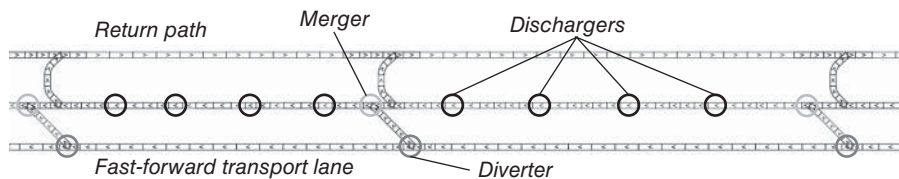


Figure 20 Area of the BHS layout with indication of diverters, mergers, and dischargers.

A diverter (in the bottom lane) has the option to detour nonurgent baggage to the middle lane to give way for urgent baggage in the transport line, but with no queues in the system all totes should follow the shortest path. When the routes merge again at the mergers in the middle lane, that lane will give higher priority to totes from the merging leg with the most urgent baggage.

Urgency is a constructed function which gives high priority to urgent totes and negative priority to totes where remaining time to departure exceeds a threshold.

$$u_j = \begin{cases} \frac{1}{t_j^2} & t_j < U_T \\ \frac{1}{(U_{\max} - U_T)^2} (-t_j^2 + 2U_T t_j - U_T^2) & t_j \geq U_T \end{cases}$$

where U_{\max} is the full window size of the allocated discharger. If the tote's remaining time exceeds this value, it should go to EBS; U_T is the threshold value, which is set to 20 min, as no tote should be considered urgent if it has more than 20 min left before the discharger closes*; and t_j is the remaining time for the j th tote. The graph is plotted in Fig. 21.

The urgency factor is converted to a scale factor for the dynamic route lengths of alternatives routes. Then the principle of simple modification of the route lengths can be used here as well:

$$s_j = \begin{cases} (1 - u_j) (1 + v_{k+1}) & u_j < 0 \text{ (nonurgent tote)} \\ (1 - u_j) (1 - v_{k+1}) & u_j \geq 0 \text{ (urgent tote)} \end{cases}$$

where v_{k+1} is the aggregated urgency value for the next decision point along the route, which is requested in a communication act (FIPA request-ref) to the divert agent. The formula secures that urgent totes will group along the shortest route (as v_{k+1} is close to 1), whereas nonurgent totes are punished along the detour. If there are no queues on the routes, the v_{k+1} is 0, and the scale factor has no effect.

The mergers in the middle lane simply give higher priority to input lanes with more urgent totes. The ratio between the aggregated urgency factors of the input lanes becomes the ratio for merging totes from the input lanes.

Saturation Management

Another important strategy is trying to avoid queues by minimizing the load on the system in critical areas. Consider slow-starting queues of cars at an intersection when the light turns green.

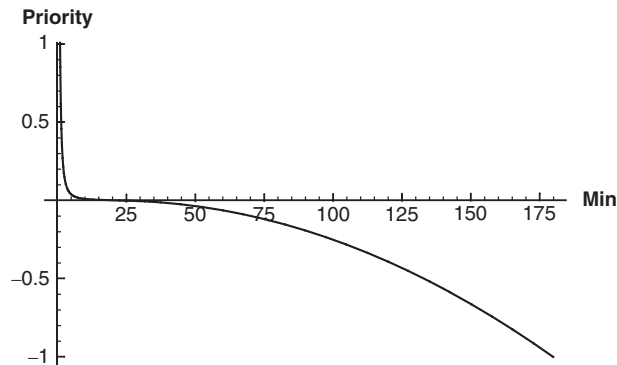


Figure 21 Urgency function for totes.

* When the discharger closes, the tote becomes a rush bag, but the threshold of 20 min is independent of the 20 min time limit for rush bags, so in total a tote is considered nonurgent if it has more than 40 min left to departure.

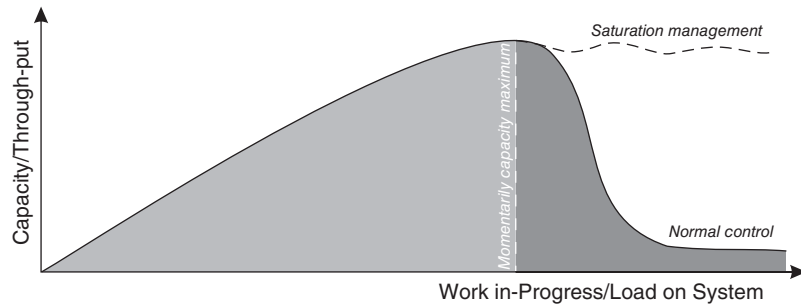


Figure 22 Theoretical WIPAC.

Acceleration ramps and reaction times relative to drivers ahead accumulate to long delays in traffic queues, even though, in theory, all drivers should be able to accelerate synchronously (no reaction time).

The same problem arises in the BHS, where reaction times correspond to the delay of the element head reporting clear.* These matters result in the characteristics of the work in progress against capacity curve (WIPAC), which is further described by Kragh,⁵⁴ who states that the capacity of a system goes down dramatically if the load on the system exceeds a certain threshold value, as indicated in the Fig. 22.

The curve is dynamical, due to the various and changing load on the system, and the maximum cannot be calculated exactly. Thus, the strategy is to quickly respond to minor observations, which indicate that the maximum has been reached, and then block new inputs to the area. This approach is called *saturation management*, where the toploders will be blocked if the routes from the toploder are overloaded.

Queues close to the toploder are most critical, as the toploders have great impact on filling up those queues, whereas the parts of the route far from the toploder could easily have been resolved before the new totes arrive. Instead of blocking the toploder completely, it can just slow down the release of new totes using the following fraction of full speed:

$$v_i = \frac{\sum_i w_i q_i}{\sum_i w_i} = \frac{\sum_i \frac{\alpha}{d_i} q_i}{\sum_i \frac{\alpha}{d_i}}$$

where v_i is the full speed of the toploder and w_i is the weight of the queue statuses, q_i , along the routes. The weight is given by a fitted coefficient, α , and the distance from the toploder, d_i . Queue statuses q_i are always a number between 0 and 1, where 1 indicates no queue.

The effect of the saturation management strategy is clearly documented by the graph in Fig. 23. Thus, the decision taken by the toploder agent is highly dependent on the current configuration of the environment around the toploder.

12 CASE STUDY 2: MATERIAL HANDLING IN AN ANODIZATION SYSTEM

The second case is a material handling system that moves bars of items between different chemical baths. Each bar has its own recipe to process the system, and system scheduling is modeled

* In the mechanical setup of the BHS, a tote can only be forwarded from one conveyor element to the next element if that element is clear. A synchronized row of totes can then pass at full speed from one element to another. In queue situations, acceleration ramps delays each element.

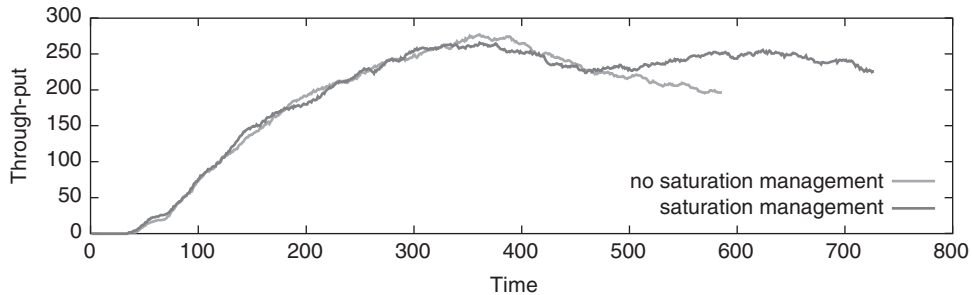


Figure 23 Result of a test scenario with and without the saturation management strategy.

by simple reactive agents that influence each other through their actions on the environment. Therefore, there is no direct negotiation between the agents; it is more a matter of coordinating their activities.

Timetabling of classes is a classical constraint satisfaction problem which is known to be hard or even NP complete for large schools or universities. But imagine the increase in complexity if duration of class sessions were allowed to vary dynamically in length. The argument could be that to take full advantage of the resources (teachers and classrooms), teachers should only stay as long as required for the students to understand the topic, but bounded by a minimum and maximum timeframe.

This case study is based on a project conducted in collaboration with Denmark's most well-known manufacturer of high-end audio and video products. The products are respected worldwide for their extremely high-quality finish and design, and the investigated production facility is the process that gives the surface of the product the high-quality finish. The process is known as an *anodization process* that increases the corrosion resistance of aluminum, but coloring of the surface is also part of the process.

In a generalized and simplified form, the problem could be described as a number of chemical baths which the items have to visit according to a prescribed recipe. Besides containing information about which baths to visit and in what order, the recipe also gives an allowed time frame for the item to stay in each bath. Items are grouped on bars with the same recipe, but a mix of different bars (that is, different recipes) could be processed at the same time in the production system.

The system consists of about 50 baths, and a typical recipe would have roughly 15–25 baths to visit. Even though all recipes do not have to visit all kind of baths, there is still room for additional baths of the same type to overcome bottlenecks, as the processing times in the bath types vary a lot. Thus, the recipe contains only bath types, not bath number, and it is the task of the control software to allocate a specific bath among duplets for every bar.

Three slightly overlapping cranes move the bars from one location to another in the array of baths. Here, a simplified notion for the movements will be used, but in practice, they are more complex than that, because moving between some specific baths includes subprocesses such as rinsing the bar of items and opening and/or closing the lid of a bath, but it comes down to an estimated travel time of moving a bar from bath u_j to bath u_{j+k} . In general, the cranes are not considered to be a bottleneck in the production system, as they handle the tasks quite sufficiently.

Apart from the baths and cranes, an important part of the system is the input buffer, where typically around 30 bars are waiting to be processed. This also is an important focus point for the control software, because choosing the best bar to fit the current configuration is the key to optimizing throughput. A general overview of the system is presented in Fig. 24, where the C

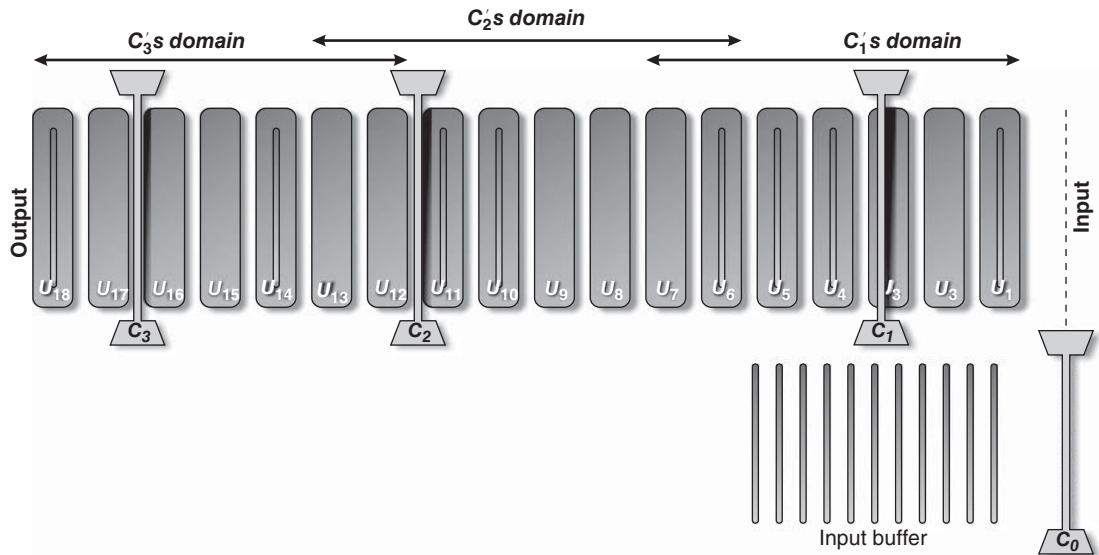


Figure 24 Generalized overview of the system for the anodization process, with bars moving from right to left.

are the cranes with a domain (how far they can move) and the U are all the chemical baths into which the bars can be placed.

At first sight, the problem described seems to be a candidate for classic optimization and scheduling principles, but as already mentioned, the allowed time frame for processing each step of the recipe complicates the task.* Another issue is the dynamical production environment, which has great impact on the system's ability to recover and finish the current bars while running as best as possible under partial breakdowns. Examples of unpredictable error conditions could be that the temperature of a bath is too low and must be heated before being available again, cranes break down, the liquid level of a bath is too low, or orders are too rapid. An agent-based approach must focus on the dynamics and be able to recover or continue as best as possible under such conditions.

The problem is classic—the throughput of a system should optimize the flow between subprocesses and handle the inflow process correctly to best utilize the system.

In abstract terms, there exists a number of tasks, q_i for $i = 1, 2, \dots, n$, with k_i subtasks[†] $q_{i,1}, q_{i,2}, \dots, q_{i,k_i}$. The subtasks are interconnected, and the order cannot be changed. Tasks should be handled as visits to processing stations—determined by the recipe.

12.1 PACO Approach

PACO is a contraction of *coordinated patterns*⁵⁵ and takes a simple approach of designing the agents. PACO focuses on reactive agents situated in an environment, where all agents are considered as partial solutions of a global problem.⁵⁶

Interactions between the agents and the environment are generally applied and modeled as forces, and by giving the agents a mass, they will—at least from a conceptual point of

* Only minimum and maximum times are given for each step of the recipe.

† Note that the number of subtasks might be different for each task group.

view—have both velocity and acceleration, which is valuable when adjusting the priorities between interactions. The applied forces are springlike forces which reduce the risk of oscillating interactions but also secure that the system will converge to an equilibrium state at some time in the coordination process between all agents.

The PACO paradigm states that agents are purely reactive; thus, they do not hold an updated internal representation of themselves, other agents, or the environment, so they have to respond to all changes of the environment in which they are situated. This general idea suits the researched case very well, as agents after an initialization process will hold some kind of plan for handling the current set of bars in the system. Whenever a new bar is introduced, or some kind of unforeseen or expected events happen within the system, such as when a crane breaks down or a bath needs cleaning, it is just a new stimulus to the agents of the control system, and they will start searching for a new equilibrium state through their interactions.

Each agent under the PACO paradigm is defined by three fields which divide the agent model in three coherent components:

1. *Perception fields* determine what the agent can perceive about its environment.
2. *Communication fields* determine with which agents an agent can interact.
3. *Action fields* determine the space in which an agent can perform its actions.

From a system point of view, the PACO paradigm also splits the system into conceptual parts, which follows the Vowels formalism explained earlier.²⁹

12.2 Agent Design

This section describes and discusses how the PACO approach under the Vowels formalism has been applied to the researched anodization system. The following sections cover each part of the method:

- *Environment.* Before agents can be created and assigned with goals and behaviors, there must be an environment for them to exist in. In this case, the environment is the baths and cranes. The environment is modeled as passive resources which the agents can ask about their status and book for a given time. Baths are accessed through a bath controller which makes baths of the same type look like only one bath in the software that is capable of containing more than one bar at a time. These baths can be asked about free space in a given direction by an agent or about whether a free time slot of a required time frame exists in a specific period. If the space is occupied, then the bath can tell which agent is blocking. A bath has no possibility to prioritize or assign time to individual agents or to push or cancel time already assigned. It is the responsibility of the agents to fight for time slots themselves.
- *Agents.* The recipe for each bar of items is split into a number of agents. One agent is created for each step of the recipe, and all agents made from the recipe form a group. An agent is born with some knowledge, as it knows which kind of bath it must visit, it holds the allowed minimum and maximum time to stay in the bath, and it knows its predecessor and successor agent of the group. It does not know the rest of the agents in the group and it has no way of communicating with them. Thus, the scope of the agent within its group is rather limited, which simplifies the interaction model.

To succeed, an agent must visit a bath of the right type, but not necessarily at the right time. The agent has a size equal to the time slot it occupies in the bath. Therefore, by its representation, agents can be seen as physical manifestations of the problem in focus. Two bars q_i

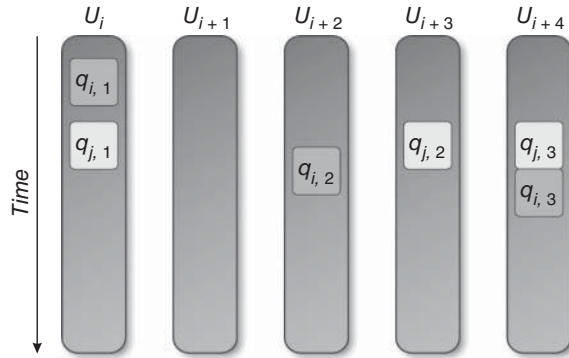


Figure 25 Three agents from each agent group occupying timeslots in the baths.

and q_j split up into two groups of agents $q_{i,1}, q_{i,2}, \dots, q_{i,1}, q_{i,2}, \dots, q_{i,n}$, and $q_{j,1}, q_{j,2}, \dots, q_{j,m}$ added to the model in random places could look like Fig. 25.

Perception, Communication, and Action Fields

As stated earlier, the PACO paradigm defines three delimited fields of how PACO agents experience the world: the perception, communication, and action fields.

- The *perception field* consists of the predecessor of the bar, as the movement of that agent affects the forces (described later in this section) applied to an agent. Furthermore, the agents above and below (in the time domain) that want to visit the same bath are also observed to avoid overlap of agents in the same bath.
- The *communication field* solely consists of the predecessor, as it should be notified if the agent could meet its goals.
- The *action field* consists of the baths of the requested type, and the organization ensures that an agent only sees one particular bath, even if the bath type is duplicated in the system.

Agent Goals

An agent has two main goals:

1. Go in the right bath.
2. Stay close to the predecessor agent of its group.

When both goals are satisfied, for all agents of a group, the bar represented by the group has a valid way to be processed by the system. Furthermore, an agent has three constraints:

1. Keep distance to both the minimum and maximum time.
2. Help the successor to stay close.
3. Help other agents in the same bath type to fulfill goals.

Constraints are added to make agents cooperate with others in fulfilling their goals, too. When agents from two groups share interests to the same time slot for a given bath, they have to be able to negotiate which will win the time slot.

Therefore, an extra type of agent is introduced—an observer. For each group of agents, one observer agent monitors the movements and how satisfied agents are in general. Information withdrawn from this observer is used when solving conflicts.

12.3 Interactions

Agents move around in the virtual world in discrete steps. They calculate a force vector v as responses to input/output from the three fields. Each discrete time step has two parts. First, all agents get a parallel chance to decide which way to go and at what speed. Hereafter, they get the chance to move themselves. In this moving step, they will try to move in the direction and distance specified by v within the space allowed by their action field.

Basic Forces

The most basic behaviors of the agents come from their primary goals and are modeled with two forces: a spring force and a gravity force. The spring force, F_s , represents the attraction to the predecessor, if any, and attracts the agent toward the point where the predecessor's time slot of the previous bath ends, so the bar can move from one bath to another, which is a criterion for a plan to be valid.

In general, a spring force is denoted $F_s = -kx$, where k is the spring constant and x the distance, so in this case $F_s = -k_{\text{parent}}(x - x_p)$, where k_{parent} is a static constant, x is the position of the agent, and x_p is the ending point of the predecessor.

The second force, the gravitational force, tries to pull the agent up. Up in the virtual model represents the beginning of time in the real world, as shown in Fig. 26. The gravitational force is given by $F_g = mg$, where m is the mass of the agent and g the gravitational acceleration. With the mass of all agents being the same, $F_g = k_g$, where k_g is a static constant force vector. This gravitational force is only applied to an agent when it is floating freely. If the agent is in contact with another, in the direction pulled by the force, the counterforce from the contact will cancel out the gravity. The total force is denoted F_t :

$$F_t = F_s + F_g$$

With only these two simple forces, a set of agents can be added that can align themselves and thereby make a valid schedule of how to be processed by the system. See Fig. 26, where U_{i+1} is the next bath the q_i bar has to visit, the F_s are the spring forces between these agents, and the F_g are the gravitational forces between these agents.

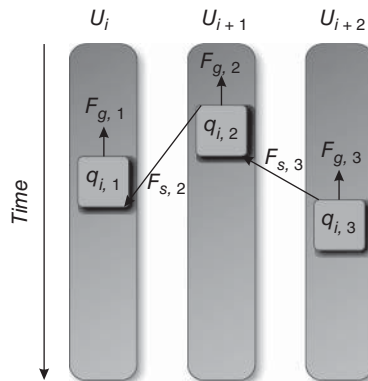


Figure 26 Basic forces for the agents.

The spring force serves to compact the plan of an agent group in order to minimize the total processing time of a bar, whereas the gravity force works to compact the entire plan for all bars in order to maximize utilization of all baths.

Organizations

To make the interaction between the agents more flexible, six social laws are introduced:

Law 1

If there is a certain amount of free space around the agent, increase size to

$$T_{\text{current}} = T_{\text{min}} + X(T_{\text{max}} - T_{\text{min}})$$

where X is a static constant and T_{current} , T_{min} , and T_{max} are the current, minimum, and maximum time slots of the agent.

Law 2

If another agent using the same bath approaches within a given distance, then shrink the current size until T_{min} plus a given margin is reached.

Law 3

If the successor is unable to reach its second goal, then stepwise increase T_{current} until T_{max} is reached.

If an agent needs to go in a direction blocked by other agents, it should be able jump over, push, or switch places with one of the blocking agents, as illustrated in Fig. 27. For this purpose, the remaining social laws apply. They are respected when agent A wants to go in a direction blocked by agent B .

Law 4

If F_t for A is greater than the current size of B and a time slot of at least A_{min} is available between the end of B and the length of F_t , A jumps to the other side of B without notice.

Law 5

If there is no room for A on the other side of B but B is trying to move in the opposite direction and if the size of F_t for A is greater than half of B_{min} , then they switch places.

Law 6

If neither of two previous laws applies but A still wants some or all of the time slot assigned to B , then A starts a negotiation based on the general satisfaction of groups A and B . If A wins this negotiation, B is pushed away; otherwise they both will have to stay.

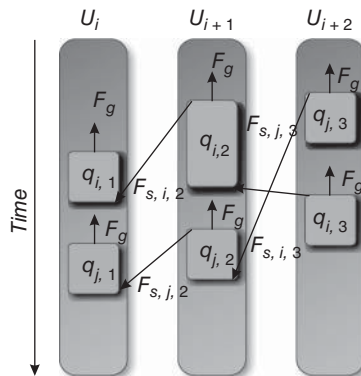


Figure 27 Conflict between two agent groups.

With this set of forces and laws for solving problems, a decent number of bars should be able to be split up into PACO agents and added to the system. Hereafter, they will align themselves in a roughly optimal way or stay in motion, trying to seek their goals.

Apparently, the real challenge of applying the PACO paradigm to an agent-based control system like the one in this case study is designing and fitting the forces used for interactions.

The agent group, which spawns from the creation of agents for a single bar of items, is not modeled or implemented as a sole entity in the system. Thus, no overall goal or intentions of the group can be directly implemented but must be realized through the aggregation of subgoals met by the agents within the group. The tension appearing inside a group due to the spring forces of the agents can, to some degree, lead to competitions among agents within a group, but the social laws make it easier for the system to reach an equilibrium and dampen the interagent tensions. Particularly, laws 1 and 3 are added to cope with these side effects of the basic forces. Law 1 simplifies the process of attraction and stabilizes the movements of a successor agent to its predecessor, due to the expansion of the current time slot for an agent in a bath, if it is too hard to pack the schedule for a bar tighter. Note that the plan for each bar at the end must form a consecutive sequence of visits to baths as the cranes move bars from bath to bath, because the system has no spare slots that temporarily can hold a bar. Law 3 more directly compacts the plan of a group and increases robustness in the coordination process.

Law 2 is important as well, even though it is orthogonal to the agent groups. It adds flexibility by minimizing the time slot requested by an agent. It is not a direct coordination mechanism between agents from different groups but allows some mutual impact on their actions.

The most challenging part of optimizing the overall plan for the system is to decide when and how conflicts between agents should be solved. No method or measurement exists to validate if a current configuration is optimal or jumps between the agents should be handled. Laws 4 and 5 direct the trivial conflicts to be handled without contracting classic local optimization principles. Law 6 serves to dampen intergroup tensions, especially to avoid oscillating shifts between agents from different groups with interest in the same bath.

13 RESULTS

The agent-based approach for the control system can be tested to see if it can create valid plans for the system, which can be done by measuring a satisfactory rate for an agent group. A fully valid plan would have 100% satisfactory rate, which means that for a given bar all visits to baths in the recipes comply with the minimum and maximum time frames and that moves between two consecutive visits are connected with no glue time (the extra time added to a bath visit for the plan to consist of consecutive visits between the different baths).

Naturally, the computation time is also interesting as a measure of the dynamic reactivity of the approach, but as the results show in Fig. 28, the agents relative quickly move to a rather stable level of their satisfaction rate.

Valid plans are not met in all scenarios, so in order to improve the results, a number of experiments have been conducted which clearly improve the number of valid plans being generated. Some of those strategies will be explained in this section. On a metalevel, they show one of the real strengths for developers to work with agent systems: It is very easy to add or change a behavior that is local in one agent and test whether the performance has improved without restructuring a central control. These strategies also give the agent a more deliberative behavior, so by introducing such changes, the system would be more a hybrid system than a system of purely reactive agents.

13.1 Active, Sleeping, and Locked Agents

During tests, it has generally been observed that the system could end in a nonconverging situation where one or more agents oscillate. Thus, a promising approach is to give the agents

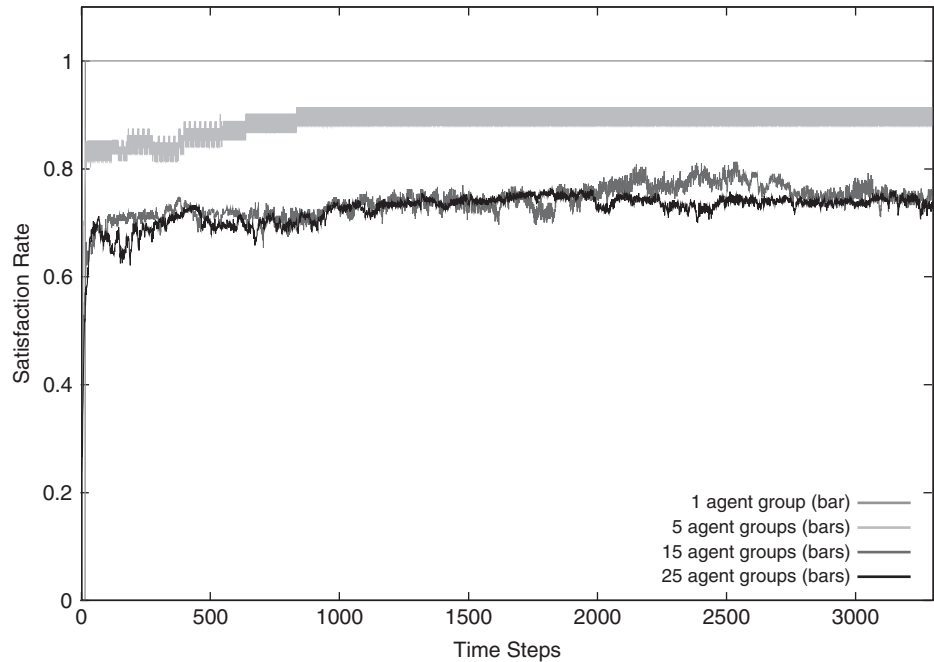


Figure 28 Satisfaction rate for test with 1, 5, 15, and 25 agent groups.

a state that determines their ability to perceive the environment and how they should react to new stimuli. Three states are obvious for the agents:

1. *Active agents* are agents that observe everything of their fields and fully accept the input and influence of other agents according to the six laws. In other words, active agents behave as expected from the design section.
2. *Sleeping agents* are agents that no longer possess intentions to move and that other agents no longer can expect to influence. However, stimuli from the environment and input from other agents can grow so strong that the agent is awakened again.
3. *Locked agents* are agents that no longer are under the influence of the environment or other agents but internally can still decide to unlock and become active again.

The interesting issues are the transitions from state to state. For a sleeping agent, there are two options: The agent itself or its group expects that the agent can improve its position or the agent is being pushed by the environment. An agent could expect to improve its position if free space above has become available and not necessarily directly above the agent. This could also happen if a larger chunk of free space has become available earlier in the time domain.

The pressure from other agents can be controlled by a threshold value, so the agent is awakened if the forces applied from other agents are too high, given by the summed force of impact from others:

$$F_{I_o} = \sum_i F_i S_i d_i$$

where F_i is the force from the i th agent that wants the position, S_i is the satisfaction rate of the i th agent, and d_i is its distance in time to the sleeping agent.

Also, the pressure on an agent from its group can be expressed as a summing force that can break a threshold value and bring the agent awake again:

$$F_{I_G} = \sum_j F_j \frac{1}{d_j}$$

where F_j is the applied force from the j th agent in the group and d_j its distance in steps to the sleeping agent. Given those transitions, an agent could fall asleep if it has found a steady state and is not in conflict with other agents. It could also fall asleep if its group has found a stable level but some members are oscillating.

13.2 Predecessor Validation

An agent adjusts its position according to the free space around it but also under the influence of its predecessor's position. Experiments have shown that, especially during the initial settling time of an agent group, some agents were strongly influenced by their predecessor agents, which were not very reliable with respect to their final position.

In the example, agent $q_{i,3}$ would move upward in time due to gravitation and the position of $q_{i,2}$, but it is rather obvious to see that $q_{i,2}$ is not very reliable due to position of $q_{i,1}$ that seems to have settled next to $q_{i,1}$. It is not certain that the parent validation will improve the plan of the system, but it is introduced to dampen oscillations of agents. One way to validate the parent is to look at its position according to its parent, as illustrated by the dimension in Fig. 29.

Figure 30 shows the result of a scenario with 25 bars both with and without the parent validation. As expected, the plan becomes more stable with the parent validation, but it also has a longer settling time as a natural consequence.

13.3 Floating

The last improvement is called *floating agents* and is best described from Fig. 31.

According to Fig. 31, an agent q_i would behave as in the left case (a). The action field of an agent only allows an agent to move in the direction of the resulting force until it is blocked by other agents, in this case q_j , even though the force is larger. By extending the action field to the size of the resulting force and allowing the agent to float over another agent q_j , many conflicts might be avoided. It is similar to allow the agent to search for a valid position from the bottom

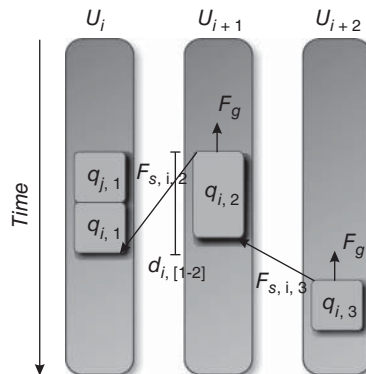


Figure 29 Parent validation problem.

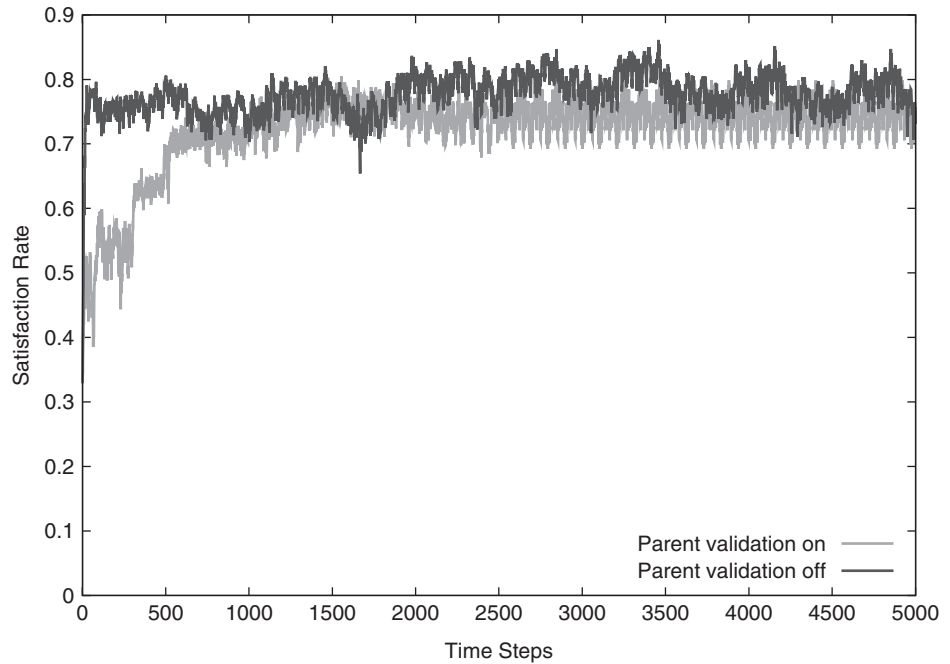


Figure 30 Satisfaction rate for 25 agent groups with and without parent validation.

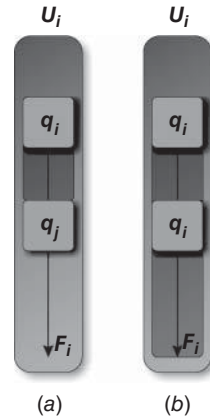


Figure 31 Floating improvement—adjusting the action field.

of its action field, as illustrated in the right case (*b*), whereas the agent in (*a*) searches from the top until it meets a block or the end of the force vector.

Figure 32 presents the results of the scenario with and without the floating improvement enabled. There might be more fluctuations with floating enabled, but as expected, it dramatically improves not only the settling time but also the overall satisfaction level, as the agents avoid many conflicts caused by tensions between agents.

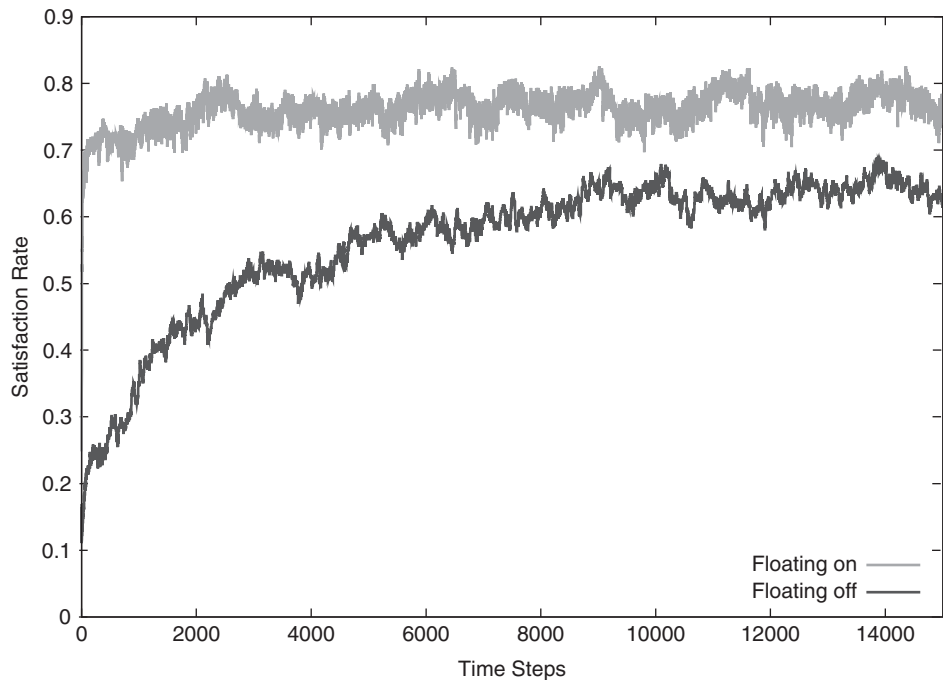


Figure 32 Satisfaction rate for 25 agent groups with and without floating enabled.

14 SUMMARY

This chapter introduced intelligent control mechanisms for manufacturing and material handling systems. First, a historical overview has shown how the technological evolution and higher demands for consumers have led to new challenges for the manufacturers.

A conceptual introduction to agent technologies was given which summarizes some of the important characteristics about both the single agent and when they collectively become a multiagent system.

Architectures which describe the internal structure of the agent were discussed in the perspective of different agent types. Reactive agents directly respond to perceived inputs, whereas deliberative agents proactively pursue their own goals.

Also, the principles for communication and collaboration for agents were presented in the chapter. Coordination and negotiation are the dominating approaches for manufacturing and material handling systems, but in most cases a control system will use a hybrid of the two approaches. The means of collaboration are strongly influenced by the organization of the agents. This chapter introduced a number of structures to facilitate this, such as a hierarchical or team-based organization.

Finally, the chapter concluded with two practical examples of different agent-based manufacturing and material handling systems that were very different in their approach for using agent technologies: a baggage handling system composed of highly deliberative agent negotiating all the actions in the system and a control system for transferring items between baths in a chemical process, where agents were simple reactive agents that coordinated their actions to reach a solid plan for the global system.

REFERENCES

1. S. Davis, "Mass Customizing," *Planning Rev.*, **17**(2), 16–21, 1989.
2. D. Sipper and R. Bulfin, *Production: Planning, Control and Integration*, McGraw-Hill College, New York, 1997.
3. F. W. Taylor, *The Principles of Scientific Management*, Harpers & Brothers Publishers, New York, 1911.
4. R. W. Brennan and D. H. Norrie, "From FMS to HMS," in S. M. Deen (ed.), *Agent-Based Manufacturing—Advances in the Holonic Approach*, Springer, Berlin, 2003, pp. 31–49.
5. D. Williamson, "System 24—A New Concept of Manufacture," in *Proceedings of the 8th International Machine Tool and Design Conference*, University of Manchester, September, Pergamon Press, Oxford, 1967, pp. 327–376.
6. P. M. Swamidass, *Manufacturing Flexibility*, Monograph No. 2, Operations Management Association, 1988.
7. N. Sandell, P. Varaiya, M. Athans, and M. Safonov, "Survey of Decentralized Control Methods for Large Scale Systems," *IEEE Trans. Automatic Control*, **AC-23**(2), 108–128, 1978.
8. S. Bussmann, "An Agent-Oriented Architecture for Holonic Manufacturing Control," in *Proceedings of First Workshop on Intelligent Manufacturing Systems Europe*, Lausanna, Switzerland, EPFL, 1998, pp. 1–12.
9. H. V. D. Parunak, *Autonomous Agent Architectures: A Non-Technical Introduction*, Industrial Technology Institute, ERIM, 1995.
10. H. V. D. Parunak, "Applications of Distributed Artificial Intelligence in Industry," in G. O'Hara and N. Jennings (Eds.), *Foundations of Distributed Artificial Intelligence*, Wiley Interscience, New York, 1996, pp. 139–163.
11. J. H. Christensen, "HMS: Initial Architecture and Standard Directions," in *Proceedings of the 1st European Conference on Holonic Manufacturing Systems*, HMS Consortium, Hannover, Germany, 1994, pp. 1–20.
12. A. Koestler, *The Ghost in the Machine*, Penguin, London, 1967.
13. R. W. Brennan and D. H. Norrie, "Agents, Holons and Function Blocks: Distributed Intelligent Control in Manufacturing," *J. Appl. Syst. Studies: Special Issues on Industrial Applications of Multi-Agent and Holonic Systems*, **2**(1), 1–19, 2001.
14. V. Mařík and M. Pěchouček, "Holons and Agents. Recent Developments and Mutual Impacts," in *Proceedings of the Twelfth International Workshop on Database and Expert Systems Applications*, IEEE Computer Society, 2001, pp. 605–607.
15. H. V. D. Parunak, "Manufacturing Experience with the Contract Net," in M. N. Huhns (Ed.), *Distributed Artificial Intelligence*, Pitman, London, 1987, pp. 285–310.
16. S. Brückner, J. Wyls, P. Peeters, and M. Kollingbaum, "Designing Agents for Manufacturing Control," in *Proceedings of the 2nd Artificial Intelligence and Manufacturing Research Planning Workshop*, August 1998, pp. 40–46.
17. S. Bussmann and K. Schild, "An Agent-Based Approach to the Control of Flexible Production Systems," in *Proceedings of 8th IEEE Int. Conf. on Emerging Technologies and Factory Automation (ETFA 2001)*, Antibes Juan-les-pins, France, 2001, pp. 169–174.
18. W. Shen and D. Norrie, "Agent-Based Systems for Intelligent Manufacturing: A State-of-the-Art Survey," *Int. J. Knowl. Inf. Syst.*, **1**(2), 129–156, 1999.
19. K. Asai, S. Takashima, and P. R. Edwards, *Manufacturing Automation Systems and CIM Factories*, Chapman and Hall, London, 1994.
20. U. Rembold and B. O. Nnaji, *Computer Integrated Manufacturing and Engineering*, Addison-Wesley, Boston, MA, 1993.
21. A.W. Colombo, "An Agent-Based Intelligent Control Platform for Industrial Holonic Manufacturing Systems," *IEEE Trans. Ind. Electron.*, **53**(1), 2006.
22. A.-W. Scheer *Architecture of Integrated Information Systems: Foundations of Enterprise Modelling*, Springer-Verlag, New York, 1992.

23. L. D. Erman and V. R. Lesser, "A Multi-Level Organization for Problem Solving Using Many, Diverse, Cooperating Sources of Knowledge. Marts," in *Proceedings of the Fourth International Joint Conference on Artificial Intelligence*, 1975, pp. 483–490.
24. R.G. Smith, "A Framework for Distributed Problem Solving," in *Proceedings of the 6th International Joint Conference of Artificial Intelligence*, 1979, pp. 836–841.
25. C. Hewitt, "Viewing Control Structure as Patterns of Passing Messages," *J. Artif. Intell.*, 8, June 1977.
26. M. Wooldridge, *An Introduction to Multiagent Systems*, Wiley, New York, 2002.
27. M. Wooldridge and N. Jennings, "Intelligent Agents: Theory and Practice," *Knowledge Eng. Rev.*, 10(2), 115–152, 1995.
28. S. Bussmann, N. R. Jennings, and M. Wooldridge, *Multiagent Systems for Manufacturing Control—A Design Methodology*, Springer, Berlin, 2004.
29. J. L. T. da Silva and Y. Demazeau, "Vowels Co-ordination Model," in *AAMAS '02: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*, ACM Press, New York, 2002, pp. 1129–1136.
30. H. V. D. Parunak, "Go to the Ant: Engineering Principles from Natural Multi-Agent Systems," *Ann. Oper. Res.*, 75, 69–101, 1997.
31. J. Ferber, O. Gutknecht, and F. Michel, "From Agents to Organizations: An Organizational View of Multi-Agent Systems," in *Proceedings of AOSE'03*, LNCS, Vol. 2935, Springer, 2003, pp. 214–230.
32. F. Zambonelli, N. Jennings, and M. Wooldridge, "Organisational Abstractions for the Analysis and Design of Multi-Agent Systems," paper presented at The First International Workshop on Agent-Oriented Software Engineering, Limerick, Ireland, 2000.
33. A. S. Rao and M. P. Georgeff, "An Abstract Architecture for Rational Agents," in C. Rich, W. Swartout, and B. Nebel (Eds.), *Proceedings of Knowledge Representation and Reasoning (KR&R-92)*, pages 439–449, 1992.
34. P. Valckenaers, P. V. Hadeli, M. Kollingbaum, H. van Brussel, and O. Bochmann, "Stigmergy in Holonic Manufacturing Systems," *Integrated Computer-Aided Eng.*, 9(3), 281–289, 2002.
35. R. A. Brooks, "A Robust Layered Control System for a Mobile Robot," *IEEE J. Robotics Automation*, 2(1), 14–23, 1986.
36. C. W. Reynolds, "Flocks, Herds, and Schools: A Distributed Behavioral Model," *Computer Graphics (SIGGRAPH '87 Conf. Proc.)*, 21, 25–34, 1987.
37. J. Muller, *The Design of Intelligent Agents: A Layered Approach*, Springer, 1996.
38. M. Wooldridge, "Intelligent Agents," in G. Weiss (Ed.), *Multi-Agent Systems*, MIT Press, Cambridge, MA, 1999, pp. 27–77.
39. T. W. Malone and K. Crowston, "The Interdisciplinary Study of Coordination," *ACM Comput. Surv.*, 26, 87–119, 1994.
40. N. R. Jennings, "Coordination Techniques for Distributed Artificial Intelligence," in G. M. P. O'Hara and N. R. Jennings (Eds.), *Foundations of Distributed Artificial Intelligence*, Wiley, New York, 1996, pp. 187–210.
41. K. Decker, *TAEMS: A Framework for Environment Centered Analysis and Design of Coordination Mechanisms*, Wiley Interscience, New York, 1996, pp. 429–448.
42. K. Decker, "Environment Centered Analysis and Design of Coordination Mechanisms," Ph.D. thesis, Department of Computer Science, University of Massachusetts, May 1995.
43. K. Decker and J. Li, "Coordinated Hospital Patient Scheduling," in *Proceedings of the Third International Conference on Multi-Agent Systems (ICMAS98)*, IEEE Computer Society, Washington, D.C., 1998, pp. 104–111.
44. K. Decker, and J. Li, "Coordinating Mutually Exclusive Resources Using GPGP," *Auton. Agents Multi-Agent Syst.*, 3(2), 133–157, 2000.
45. G. Maionea and D. Naso, "A Soft Computing Approach for Task Contracting in Multi-Agent Manufacturing Control," *Comput. Ind.*, 52(3), 199–219, 1996.
46. A. Giret, and V. Botti, "Analysis and Design of Holonic Manufacturing Systems," in *Proceedings of 18th International Conference on Production Research (ICPR2005)*, 2005.
47. D. G. Pruitt, *Negotiation Behaviour*, Academic, New York, 1981.

48. T. Finin, J. Weber, G. Wiederhold, M. Genesereth, R. Fritzon, D. McKay, J. McGuire, R. Pelavin, S. Shapiro, and C. Beck, *Specification of the KQML Agent-Communication Language*, The DARPA Knowledge Sharing Initiative External Interfaces Working Group, 1993.
49. J. R. Searle, *Speech Acts*, Cambridge University Press, London, 1969.
50. B. Hurling and V. Lesser, "Using Quantitative Models to Search for Appropriate Organizational Designs," *Autonom. Agents Multi-Agent Sys*, **16**(2), 95–149. 2008.
51. A. J. M. Donaldson, "A Case Narrative of the Project Problems with the Denver Airport Baggage Handling System (DABHS)," TR 2002–01, Middlesex University, School of Computing Science, January 2002.
52. E. W. Dijkstra, "A Note to Two Problems in Connexion with Graphs," *Numer. Math.*, **1**, 269–271, 1959.
53. M. Negnevitsky, *Artificial Intelligence—A Guide to Intelligent Systems*, 2nd ed., Addison Wesley, Reading, MA, 2005.
54. A. Kragh, "Kø-Netværksmodeller til Analyse af FMS Anlæg," Ph.D. Thesis, Informatics and Mathematical Modelling, Technical University of Denmark, 1990.
55. Y. Demazeau, "Coordination Patterns in Multi-Agent Worlds: Application to Robotics and Computer Vision," in *IEEE Colloquium on Intelligent Agents*, IEEE, London, February 1991.
56. Y. Gufflet and Y. Demazeau, "Applying the PACO Paradigm to a Three-Dimensional Artistic Creation," in *Proceedings of 5th International Workshop on Agent-Based Simulation (ABS'04)*, May 2005, Lisbon, Portugal, 2004, pp. 121–126.
57. Y. Demazeau, "From Interactions to Collective Behaviour in Agent-Based Systems," in *Proceedings of the First European Conference on Cognitive Science*, Saint Malo, France. 1995, pp. 117–132.
58. G. Di Caro and M. Dorigo, "AntNet: A Mobile Agents Approach to Adaptive Routing," IRIDIA/97-12, Université Libre de Bruxelles, Belgium, 1997.
59. "FIPA Abstract Architecture Specification," SC00001L, Foundation for Intelligent Physical Agents, December 2002.
60. "FIPA Communicative Act Library Specification," SC00037J, Foundation for Intelligent Physical Agents, December 2002.
61. C. Flake, C. Geiger, G. Lehrenfeld, W. Mueller, and V. Paelke, "Agent-Based Modeling for Holonic Manufacturing Systems with Fuzzy Control," in *Proceedings of 18th International Conference of the North American Fuzzy Information Processing Society (NAFIPS'99)*, New York, 1999, pp. 273–277.
62. T. W. Malone, K. Crowston, J. Lee, B. Pentland, C. Dellarocas, G. Wyner, J. Quimby, C. Osborne, A. Bernstein, G. Herman, M. Klein, and E. O'Donnel, "Tools for Inventing Organizations: Towards a Handbook of Organizational Processes," *Manag. Sci.*, **45**(3), 425–443. 1999.
63. V. Mařík, M. Pěchouček, P. Vrba, and V. Hrdonka, "FIPA Standards and Holonic Manufacturing," in S. M. Deen (Ed.), *Agent-Based Manufacturing—Advances in the Holonic Approach*, Springer, Berlin, 2003, pp. 31–49.
64. H. V. D. Parunak, "Industrial and Practical Applications of DAI," in G. Weiss (Ed.), *Multiagent Systems—A Modern Approach to Distributed Artificial Intelligence*, MIT Press, Cambridge, MA, 1999, pp. 377–421.
65. R. Schoonderwoerd, O. Holland, J. Bruten, and L. Rothkrantz, "Ant-Based Load Balancing in Telecommunication Networks," *Adaptive Behav.*, **5**, 169–207, 1996.

CHAPTER 18

MANAGING PEOPLE IN ENGINEERING AND TECHNOLOGY

Hans J. Thamhain
Bentley University
Waltham, Massachusetts

1 CRITICAL ISSUES AND CHALLENGES	559	4 POWER PROFILE IN ENGINEERING AND TECHNOLOGY MANAGEMENT	569
2 UNIQUE NATURE OF MANAGING ENGINEERING PERSONNEL	561	5 MANAGING AND LEADING ENGINEERING TEAMS	570
3 MOTIVATION AND ENGINEERING PERFORMANCE	564	5.1 Building High-Performance Teams	572
3.1 Implications for Engineering Performance	566	6 RECOMMENDATIONS FOR EFFECTIVE ENGINEERING MANAGEMENT	573
3.2 Motivation as Function of Risks and Challenges	566	7 CONCLUDING REMARKS	577
3.3 Managing in Range of High Motivation	567	REFERENCES	577

1 CRITICAL ISSUES AND CHALLENGES

People are the most critical resource driving business performance. Business leaders and scholars agree.^{1–8} Yet, managing these people is very difficult. It is especially challenging in today's complex business environment with many operations distributed across the globe. This requires working with people from different support organizations, vendors, partners, customers, and government agencies; it requires effective networking and cooperation among organizations with different cultures, values, and languages.¹⁰ Thus, managers today, especially in engineering and technology, must be capable of dealing with the technical challenges as well as the economic, political, social, and regulatory issues and the associated uncertainties and risks,^{16–19} as profiled in Table 1.

All of this has an impact on managerial education, training, and skill development. Not too long ago, managers *could* successfully lead their work groups toward desired results by focusing on the work requirements, timing, and resource constraints. However, these traditional approaches are no longer sufficient. They have become threshold competencies, critically important, but unlikely to guarantee success by themselves.

The mandate for managers is clear: They must weave together the best practices and programs for continuously developing their people toward highest possible performance. However, even the best practices and most sophisticated methods do not guarantee success.

Table 1 Core Management Issues and Challenges

Manage Technical Work Content. Any job has technical content. But, especially in engineering and technology, assignments and projects are technologically intense and complex. The ability to manage the technical work, on an individual job level and collectively throughout the organization, is a threshold competency critically important to the success of any technology-based business. Managerial issues relate to staffing, professional development, support technologies, innovation, and risk management.

Manage Talent. While equipment, buildings, and infrastructure are important, businesses succeed because of their people, their ideas, and the actions that bring the system alive. For many technology companies, talent is everything! The type of talent and its fit with the business needs and organizational culture determine everything from idea generation to problem resolution and business results. Talent does not occur at random. Nor should it be taken as granted. It needs to be searched out, attracted, developed, and maintained. An organization's personal policies and award systems must be consistent with the talent objectives. Losing a top talent is a sin! Companies like GE conduct postmortems on every top talent loss and hold their management accountable for those losses.

Manage Knowledge. Technology companies are knowledge factories. In essence, they buy, trade, transfer, and sell knowledge. Their value lies increasingly in the collective knowledge that becomes the basis for creating new ideas, concepts, products, and services. The emphasis must be on orchestrated management of this collective multidisciplinary knowledge. New products and services usually do not come from a single brilliant idea but are the result of broad-based collaborative efforts throughout the organization. They are the result of an intricately connected, vast knowledge network with high interconnectivity and low cross-organizational impedance. Setting up effective support systems and managing the development, processing, filtering, sharing, and transferring knowledge toward desired results are very important and challenging tasks which require sophisticated people skills.

Manage Information. Similar to knowledge, information management has a strong human side which often does not receive sufficient attention. Regardless of the available technology, people are involved in gathering, transferring, interpreting, sharing, and acting upon information.

Manage Innovation and Creativity. The important role of technological innovation for business success has been long recognized.²⁰ Innovation can generate competitive advantages for one firm while eroding the market position for another, which is especially true for enterprises that derive their added value from technology. "Our technology, reach and resources aren't enough to make us the global best. It's all about people, nurturing, energizing, and inspiring them to search for ideas and to cooperate. It's about creating a culture that brings everyone into the game across the organization." This statement of GE's CEO, Jeff R. Immelt, in an address to the company's shareholders is typical for the critical role of innovation in today's ultra-competitive world of business. However, generating such a competitive advantage from innovation is an intricate and risky process. The companies that will survive and prosper in the decades ahead will be those that can manage innovation and derive business benefits from it. They must do this in spite of the complex organizational processes, increasing risks, uncertainties, and rapidly changing markets and technology.

Manage Communications. Communication is the backbone of a firm's command-and-control structure. It is the catalyst toward crucial integration of organizational efforts toward unified results. This is especially critical in technology firms with their unconventional organizational structures and strong need for cross-functional coordination. While strongly supported by IT, communication systems are much broader in scope. They include everything from digitized systems to focus groups, task action teams, and old-fashioned group meetings.

Build Supportive Organizational Environment. Successful companies have cultures and environments that support their people. These companies provide visibility and recognition to their people. They also show the impact of these accomplishments on the company's mission and project-related objectives. This creates an ambiance where people are interested and excited about their work, which produces higher levels of ownership, cross-functional communications, collaboration, and commitment. These are also the conditions for unifying team members behind the requirements and keeping the project effort focused and synchronized.

Ensure Direction and Leadership. Managers themselves are change agents. Their concern for people, assistance to problem solving, and enthusiasm for the enterprise mission and objectives can foster a climate of high motivation, work involvement, commitment, open communications, and willingness to cooperate across the organization.

They must be carefully integrated with the business process, its culture, and its value system. The challenges are especially felt in today's technology organizations, which have become highly complex internally and externally, with a bewildering array of multifaceted activities, requiring sophisticated cross-functional cooperation, integration, and joint decision making. Because of these dynamics, technology organizations are seldom structured along traditional functional lines; rather they operate as matrices or hybrid organizations with a great deal of power and resource sharing. In addition, lines of authority and responsibility blur among formal management functions, project personnel, and other subject experts, leading to a more empowered and self-directed work force with a much higher bandwidth of skills to solve operational problems responsive to the needs of our ultracompetitive business environment.

New technologies, especially computers and communications, have radically changed the workplace and transformed our global economy, focusing on effectiveness, value, and speed. These new technologies offer advanced capabilities for cross-functional integration, resource mobility, effectiveness, and market responsiveness, but they also require more sophisticated skill sets both technically and socially, dealing effectively with a broad spectrum of contemporary challenges, including managing conflict, change, risks, and uncertainty. As a result of this paradigm shift, the dynamics of how people work together has changed, and managerial focus has shifted from efficiency to effectiveness and from traditional performance measures, such as the quadruple constraint (deliverables, budget, and schedule), to include a broader spectrum of critical success factors that support process integration effectiveness, organizational collaboration, innovation, human factors, and overall business process effectiveness and strategic objectives. These issues involve great challenges which define a new management frontier as summarized in Table 1.

2 UNIQUE NATURE OF MANAGING ENGINEERING PERSONNEL

Managers in engineering and technology see themselves different from those in less technical environments. Their work requires unique organizational structures, policies, and interactions among people. Their management style and leadership must not only be consistent with the nature of the work and the business process but also be conducive to the special needs of the people and consistent with the unique culture of the technology-based organization and its values.

Responding to these special needs, many engineering organizations evolved over time with their own somewhat unique characteristics, making engineering and technology management a specific discipline, different from other types of management. Figure 1 shows six of the most influential organizational subsystems— (1) *people*, (2) *work*, (3) *business process*, (4) *organizational culture*, (5) *leadership*, and (6) *external business environment* —all affecting the work environment, team characteristics, and ultimately enterprise performance, as discussed below.

- *People (Skill Sets, Traits, and Attitudes)*. Because of the type of work and challenges, engineering-oriented environments attract different people, with specialized knowledge and skill sets. Usually, they are better educated and self-motivated and require minimum supervision. They enjoy problem solving and find technical challenges motivating and intellectually stimulating. They enjoy a sense of community and team spirit, while having low tolerance for personal conflict and organizational politics.

Managerial Impact. Because of the complexity of the work and the intricate decision-making processes, these work groups rely to a considerable extent on member-generated norms and performance evaluations, rather than on hierarchical guidelines, policies, and procedures.^{8,21–24} As a result, decision-making power and responsibility for achieving specific outcomes are more distributed among team members who function more autonomously. These self-directed work group models have become increasingly popular, especially for orchestrating and controlling complex projects.²⁵

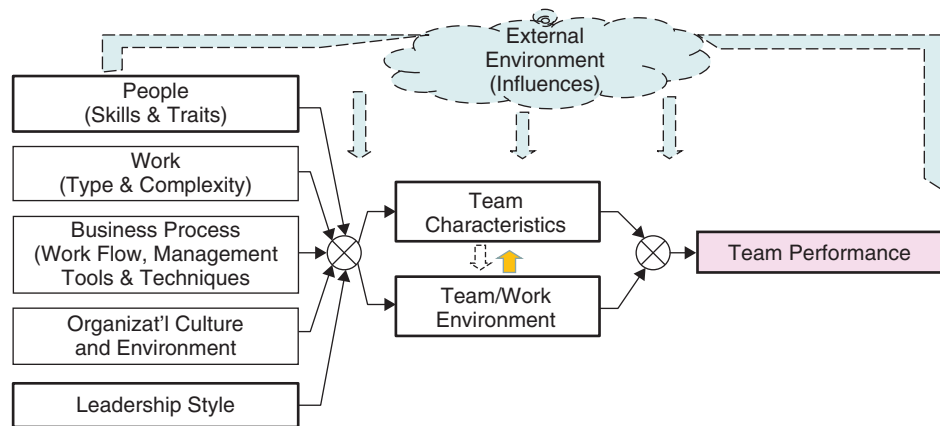


Figure 1 Influences affecting work group characteristics and performance in technology-intensive organizations.

As these contemporary teams replace traditional, hierarchical work groups, effective leadership requires a more sophisticated management style which relies strongly on group interaction, resource and power sharing, individual accountability, commitment, conflict handling, cross-functional linkages and cooperation, technology transfer models, top management involvement, and design/build approaches.²⁶ For example, traditional management tools such as static project plans and linear performance measures—designed largely for conventional management systems with clearly defined horizontal and vertical lines of communication and centralized command-and-control system—are no longer effective in these contemporary situations.^{27,28} They are often being replaced with more team-based and agile management processes, ranging from concurrent engineering to stage reviews and spiral processes. Many of these management systems, tools, and techniques are under continuous review and adjustment to adapt to the changing nature of the business environment, dynamically impacting organizational structure, work process, personnel recruiting and advancement, skill development, management style, and organizational culture.

- *Work (Type, Nature, and Complexity)*. The nature of engineering work is complex and outcomes cannot always be predicted or achieved with certainty. When describing their work, managers point to technical difficulties, high levels of innovation and creativity, evolving solutions, complex decision processes, uncertainty, and highly sophisticated forms of work integration. Much of engineering work is conducted in *projects* organized and executed by a multidisciplinary team. This involves additional challenges of resource and power sharing, intricate technology transfer networks, complex support systems, and highly sophisticated forms of performance measurements.^{16,17,29–31}

Managerial Impact. Because of the complexity and multidisciplinary nature of work, engineering organizations look for a broad talent pool, often reaching across time zones and multinational borders to form joint ventures, consortia, and partnerships. This leads to more organizational complexity and work process dynamics, which in turn require more sophisticated managerial skills to facilitate work integration, cross-functional collaboration, and control of the work toward desired results.

- *Business/Work Process (Work Flow, Managerial Tools, and Techniques)*. The way work is organized, flows through the organization, and connects with its support systems and

the technologies used for supporting the engineering work, facilitating interdisciplinary communications, and integrating work components all affect the team dynamics and effectiveness of the management style. To illustrate, a commercial airplane development results in very different organizational interactions than a pharmaceutical project with multinational R&D partners.³² A matrix-organized microprocessor rollout results in different work processes than a projectized electric car development, just to give a few examples. While many of the managerial processes, tools, and techniques are also used in other environments, the unique nature of engineering requires special tools and applications. Spiral planning, stakeholder mapping, concurrent engineering, and integrated product developments are just a few examples of the specialized nature of tools needed to produce team-based solutions to complex technical problems.

Managerial Impact. The *work process design* directly impacts people issues, management style, and organizational culture. New organizational models and management methods, such as *stage-gate*, *concurrent engineering*, and *design-build processes*, evolved together with the refinement of long-time established concepts such as the *matrix*, *project*, and *product management*. Furthermore, *managerial tools and techniques* affect the people and the work process. Matching organizational culture with any of these tools is a great challenge. Stakeholder involvement during tool selection, development, and implementation and trade-offs among efficiency, speed, control, flexibility, and risk are critical to the effective use of these tools and techniques.

- *Organizational Culture.* The challenges of technology-driven environments create a unique organizational culture with their own norms, values, and work ethics. These cultures are more team oriented regarding decision making, work flow, performance evaluation, and work group management. Authority must often be earned and emerges within the work group as a result of credibility, trust, and respect, rather than organizational status and position. Rewards come to a considerable degree from satisfaction with the work and its surroundings, with recognition of accomplishments as important motivational factors for stimulating enthusiasm, cooperation, and innovation.

Managerial Impact. Organizational culture has a strong influence on the people and work process.^{33,34} This culture is deeply rooted in the organizational fiber with strong influences on a wide range of managerial processes, affecting everything from hiring practices, performance evaluations, and reward systems to organizational structures and management style.

- *Managerial Leadership.* This is one of the strongest influences on the people and their performance in technology-based work environments. While technical skill sets, management tools, and effective work processes are absolutely critical, it is managerial leadership that drives team performance. It is the force that guides the work process, unifies the team, and fosters a culture of collaboration and commitment across organizational boundaries that connect support functions, suppliers, customers, and partners.^{1,8,35,36}

Managerial Impact. Among the many influences, managerial leadership has the strongest impact on team effectiveness and overall project success. It is also under the direct control of the manager! A better understanding of the criteria and organizational dynamics that motivate people and drive team performance can help managers in effectively integrating work groups with the enterprise. Effective team leaders are social architects who understand the interaction of organizational and behavioral variables and can foster a climate of active participation, accountability, and result orientation throughout the enterprise and its external partners. This requires an in-depth understanding of the business environment as well as its dynamics and cultures, plus sophisticated management skills in support of the credibility, trust, and respect needed for effective leadership.

- *External Business Environment.* All five previously discussed enterprise subsystems operate within a socially, politically, and economically complex business environment. Especially for engineering and technology-intensive organizations, this environment is fast changing regarding market structure, suppliers, and regulations. Short product life cycles, intense global competition, low brand loyalty, low barriers to entry, and strong dependency on other technologies and support system are very common. These complexities call for specialized work processes, new concepts of technology transfer, and more sophisticated management skills and leadership.

Managerial Impact. The need for speed, agility, and efficiency has an impact on the work process design, supply chain, organizational structure, management methods, tools, and techniques. It also affects business strategy³⁷ and competitive behavior,^{38,39} which often leads to collaboration and resource sharing via alliances, mergers, acquisitions, consortia, and joint ventures.

Engineering managers often describe their organizational environments as “unorthodox,” with ambiguous authority and responsibility relations. They argue that such environments require broader management skill sets and more sophisticated leadership than traditional business situations. In this more open, dynamic business environment, enterprise performance is based to a large degree on teamwork. Yet, attention to individuals, their competence, accountability, commitment, and sense for self-direction are crucial for organizations to function effectively. Such a “team-centered” management style is based on the thorough understanding of the motivational forces and their interaction with the enterprise environment.

3 MOTIVATION AND ENGINEERING PERFORMANCE

Understanding people, their wants and needs, is important in any management situation. It is especially critical in today’s technology-based engineering organizations. Leaders who succeed within these often unstructured environments must work with cross-functional groups and gain services from personnel not reporting directly to them. They have to deal with line departments, staff groups, team members, clients, and senior management, each having different cultures, interests, expectations, and charters. Transforming these task multidisciplinary groups into cohesive teams is difficult. To get results, these engineering managers must relate socially as well as technically and understand the culture and value system of the organization in which they work. The days of the manager who gets by with only technical expertise or pure administrative skills are gone.

What work best? Observations of best-in-class practices show consistently and measurably two important characteristics of high performers in technology organizations: (1) they enjoy work and are excited about the contributions they make to their company and society and (2) they have their professional and personal needs fulfilled. Specifically, field research studies have identified 16 of these needs that associate particularly strongly with job performance.⁴⁰

Sixteen professional needs are strongly associated with engineering job performance. As summarized in Table 2, the fulfillment of these needs drives professional people to higher performance; conversely, the inability to fulfill these needs may become a barrier to individual performance and teamwork.⁴¹ The rationale for this important correlation is found in the complex interaction of organizational and behavioral elements. Effective team management involves three primary components: (1) people skills, (2) organizational structure, and (3) management style. All three components are influenced by the specific task to be performed and the surrounding environment. That is, the degree of satisfaction of any of the needs is a function of (1) having the right mix of people with appropriate skills and traits, (2) organizing the people and resources according to the tasks to be performed, and (3) adopting the right leadership style.

Table 2 Sixteen Professional Needs Affecting Individual Team Performance

-
1. *Interesting and challenging work*: This is an intrinsic motivator that satisfies professional esteem needs and helps to integrate personal goals with organizational objectives.
 2. *Professionally stimulating work environment*: An ambience conducive to professional involvement, creativity, and interdisciplinary support. It also fosters team building and is conducive to effective communication, conflict resolution, collaboration, and commitment toward the organizational goals. The quality of the work environment is defined primarily by the organizational structure, its facilities, and its management style.
 3. *Professional growth* is measured by promotional opportunities, salary advances, the learning of new skills and techniques, and professional recognition. A particular challenge exists for management in limited-growth or zero-growth businesses to compensate for lack of promotional opportunities by offering more intrinsic professional growth in terms of job satisfaction, esteem, and skill building.
 4. *Overall leadership* involves dealing effectively with individual contributors, managers, and support personnel within a specific functional discipline as well as across organizational lines. It includes technical expertise, information-processing skills, effective communication, and decision-making skills. Taken together, leadership means satisfying the need for clear direction and unified guidance toward established objectives.
 5. *Tangible rewards* include salary increases, bonuses, and incentives as well as promotions, recognition, better offices, and educational opportunities. While financial rewards in this list are *extrinsic*, they are important and necessary to sustain strong long-term efforts, motivation, and commitment. Furthermore, they validate the “softer” *intrinsic* rewards, such as recognition and praise, and reassure people of their value in the organization.
 6. *Technical expertise* means that personnel within their work team have all necessary interdisciplinary skills and expertise available to perform the required tasks. Technical expertise includes understanding the technicalities of the work; the technology of underlying concepts, theories, and principles; design methods and techniques; and functioning and interrelationship of the various components that make up the total system.
 7. *Assisting in problem solving* includes facilitating and assisting in resolving technical, administrative, and personal issues. It is a very important need, which, if not satisfied, often leads to frustration, conflict, unsustainable commitment, and poor-quality work.
 8. *Clearly defined objectives*: Goals, objectives, and expected outcomes of an effort must be clearly communicated to all affected personnel. Conflict can develop over ambiguities or missing information.
 9. *Management control* is important for effective team performance. Managers must understand the interaction of organizational and behavior variables in order to exert the direction, leadership, and control required to steer the work effort toward established organizational goals without stifling innovation and creativity.
 10. *Job security* is one of the very fundamental needs that must be satisfied before people consider higher order growth needs. While many of the factors that influence job security are not under the control of engineering managers, through their leadership style and effective communications managers can create more or less of a favorable image and therefore affect people’s attitude and their way of dealing with a given situation.
 11. *Senior management support* should be provided in four major areas: (1) financial resources; (2) administrative support, including an effective operating charter, policies, and procedures; (3) work support from functional resource groups; and (4) necessary facilities and equipment. Management support is particularly crucial for larger, more complex undertakings.
 12. *Good interpersonal relations* are required especially for effective teamwork; they foster a stimulating work environment with low conflict, high involvement, motivated personnel, and high productivity.
 13. *Proper planning* is absolutely essential for the successful management of multidisciplinary activities. It requires communications and information-processing skills to define the actual resource requirements and administrative support necessary. It also requires the ability to negotiate for resources and commitment from key personnel in various support groups across organizational lines.
 14. *Clear role definition* helps to minimize role conflict and power struggles among team members and/or supporting organizations. Clear charters, plans, and good management direction are some of the powerful tools used to facilitate clear role definition.
 15. *Open communication* satisfies the need for a free flow of information both horizontally and vertically, keeping personnel informed and functioning as a pervasive integrator of the overall project effort.
 16. *Minimum changes*: Although technology managers have to live with constant change, their team members often see change as an unnecessary condition that impedes their creativity and timely performance. Advanced planning and proper communication can help to minimize changes and lessen their negative impact.
-

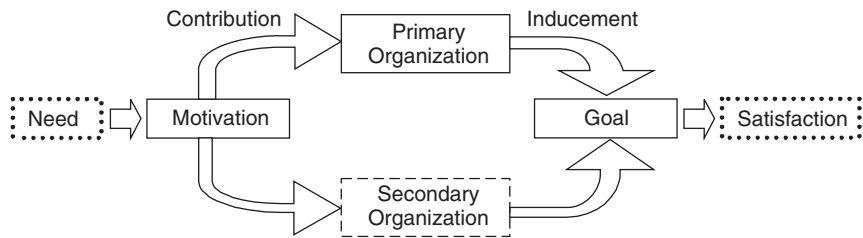


Figure 2 Inducement induction model of motivation.

3.1 Implications for Engineering Performance

The *significance* of assessing these motivational forces lies in several areas. First, Table 2 provides insight into the broad needs of engineering professionals. These needs must be satisfied *continuously* before the people can reach high levels of performance. This is consistent with findings from other studies, which show that in technology-based environments a significant correlation exists between professional satisfaction and organizational performance.² From the above listing we know now more specifically in what areas we should focus our attention. In fact, the table provides a model for *benchmarking*; that is, it offers managers a framework for monitoring, defining, and assessing the needs of their people in specific ways and ultimately building a work environment that is responsive to their needs and ultimately conducive to high performance.

Figure 2 provides further insight into the motivational process and the important role of needs via the *inducement contribution model*. The model shows that people are motivated to work toward personal and professional goals because they satisfy specific needs, such as needs for recognition, promotion, or pay increase. In the process of satisfying their needs, they make a contribution to the organization and its performance. If, on the other hand, employees cannot reach their goals and fulfill their needs through activities within their “primary organization” (i.e., employer), they may try to satisfy their needs by working toward their goals through an “outside” organization. They may involve themselves in volunteer work or recreational activities. While this may help to satisfy the employee’s need for interesting and stimulating activities, their efforts and energy are directed within a secondary organization, which provides inducement and receives his or her contributions, which is being “leaked away” from the employer.

An employee’s need satisfaction is as important to the healthy functioning of the organization as it is to the employee. Good managers are closely involved with the staff and their work. The manager should also identify any *unrealistic* goals and correct them, change situations that impede the attainment of realistic goals, and support people in reaching their goals and cheer them on, which is recognition. This will help to refuel the individual desire to reach the goal and keep the employees’ energies channeled through the primary organization. The tools that help the manager to facilitate professional satisfaction are work sign-on, delegation, career counseling and development, job training and skill development, and effective managerial direction and leadership with proper recognition and visibility of individual and team accomplishments.

3.2 Motivation as Function of Risks and Challenges

Additional insight into motivational drive and its dynamics can be gained by considering motivational strength as a function of the probability and desire to achieve the goal. We can push others, or ourselves, toward success or failure because of a mental predisposition called *self-fulfillment prophesy*. Our motivational drive and personal efforts increase or decrease

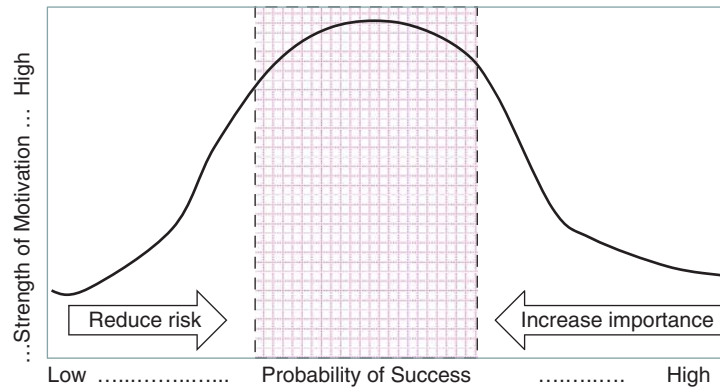


Figure 3 Pygmalion effect, relating strength to the probability of stress.

relative to the likelihood of the expected outcome. Personal motivation toward reaching a goal changes with the probability of success (perception of doability) and challenge.³⁵ Figure 3 expresses the relationship graphically. A person's motivation is very low if the probability of achieving the goal is very low or zero. As the probability of reaching the goal increases, so does the motivational strength. However, this increase continues only up to a certain level; when success is more or less assured, motivation decreases. This is an area where the work is often perceived as routine, uninteresting, and holding little potential for professional growth.

Success as a Function of Attitude. The saying “The harder you work, the luckier you get” expresses the effect of motivation in pragmatic terms. People who have a “can-do” attitude, who are confident and motivated, are more likely to succeed in their mission. Winning is in the attitude. This is the essence of the self-fulfillment prophecy. In fact, high-performing engineering teams often have a very positive image of their capabilities. As a result, they are more determined to produce the desired results, often against high risks and odds.

3.3 Managing in Range of High Motivation

Applying motivational models to the high-technology workplace suggests that managers should ensure proper matching of people to their jobs. Further, managers must foster a work environment and direct their personnel in a manner that promotes “can-do” attitudes and facilitates continuous assistance and guidance toward successful task completion. The process involves four primary issues: (1) work assignments, (2) team organization, (3) skill development, and (4) management. *Specific suggestions for stimulation and sustaining motivation in technology-oriented professionals are summarized below:*

1. Work Assignments

- Explain the assignment, its importance to the company, and the type of contributions expected.
- Understand the employee's professional interests, desires, and ambitions and try to accommodate to them.
- Understand the employee's limitations, anxieties, and fears. Often these factors are unjustly perceived and can be removed in a face-to-face discussion.
- Develop the employee's interest in an assignment by pointing out its importance to the company and possible benefits to the employee.

- Assure assistance where needed and share risks.
 - Show how to be successful. Develop a can-do attitude.
 - If possible, involve the employee in the definition phase of the work assignment, for instance, via up-front planning, a feasibility analysis, needs assessment, or a bid proposal.
2. Team Organization
- Select the team members for each task or project carefully, assuring the necessary support skills and interpersonal compatibility.
 - Select the team members for each task or project carefully, assuring the necessary support skills and interpersonal compatibility.
 - Plan each engineering project properly to assure clear directions, objectives, and task charters.
 - Assure leadership within each task group.
 - Sign-on key personnel on a one-on-one basis according to the guidelines discussed in item number 1.
3. Skill Development
- Plan the capabilities needed in your engineering department for the long range. Direct your staffing and development activities accordingly.
 - Encourage people to keep abreast in their professional field.
 - Provide for on-the-job experimental training via selected work assignments and managerial guidance.
 - Provide the opportunity for some formal training via seminars, courses, conferences, and professional society activities.
 - Use career counseling sessions and performance reviews to help in guiding skill development and matching them with personal and organizational objectives.
4. Management
- Develop interest in the work itself by showing its importance to the company and the potential for professional rewards and growth.
 - Promote project visibility, team spirit, and upper management involvement.
 - Assign technically and managerially competent task leaders for each team and provide top-down leadership for each project and for the engineering function as a whole.
 - Manage the quality of the work via regular task reviews and by staying involved with the project team without infringing on their autonomy and accountability.
 - Plan your projects up front. Conduct a feasibility study and requirements analysis first. Assure the involvement of the key players during these early phases.
 - Break activities or projects into phases and define measurable milestones with specific results. Involve personnel in the definition phase. Obtain their commitment.
 - Try to detect and correct technical problems early in their development.
 - Foster a professionally stimulating work environment.
 - Unify the task team behind the overall objectives. Stimulate the sense of belonging and mutual interdependence.
 - Refuel the commitment and interest in the work by recognizing accomplishments frequently.

- Assist in problem solving and group decision making.
- Provide the proper resources.
- Keep the visibility and priority for the project high. No interruptions.
- Avoid threats. Deal with fear, anxieties, mistrust, and conflicts.
- Facilitate skill development and technical competency.
- Manage and lead.

4 POWER PROFILE IN ENGINEERING AND TECHNOLOGY MANAGEMENT

As organizations become flatter, leaner, more agile, and self-directed, they share to a greater extent the responsibilities, resources, and power. Especially for engineering enterprises which rely to an increasing extent on innovation, cross-functional teamwork, intricate multicompany alliances, and complex forms of work integration, success depends to a considerable extent on member-generated performance norms and work processes. Self-directed team concepts are gradually replacing the traditional, more hierarchically structured organization,^{42–44} requiring a radical departure from traditional management practices of top-down, centralized command, control, and communications. To be effective, engineering managers have to direct their personnel and obtain cross-functional support without much organizationally derived power. They must develop, or “earn,” their own bases of influence and build their own power spectrum, which derives from personal knowledge, expertise, and the image of a sound decision maker. The basic concept of *power and authority* has been known for a long time. Four decades ago, French and Raven⁴⁵ presented a typology that included five bases of interpersonal power—*authority*, *reward*, *punishment*, *expertise*, and *referent power* (i.e., friendship, charisma, empathy), which are summarized in Fig. 4. To this day, these are still the most commonly recognized influences of managerial power. For some time, the first three bases—*authority*, *reward*, and *punishment*—were perceived as being derived entirely from the organization. However, more recent studies provide measurable evidence that all bases of power can be individually developed, at least to some degree. Today’s organizations grant power to their leaders in many forms. Some of it is still derived from the organizational construct and vested in the leader via organizational position, status, and other traditional components of legitimacy, including the power to reward and punishment. However, contemporary engineering managers must *earn* most of their authority and influence bases for managing their multifunctional teams. Since earned authority depends largely on the image of trust,

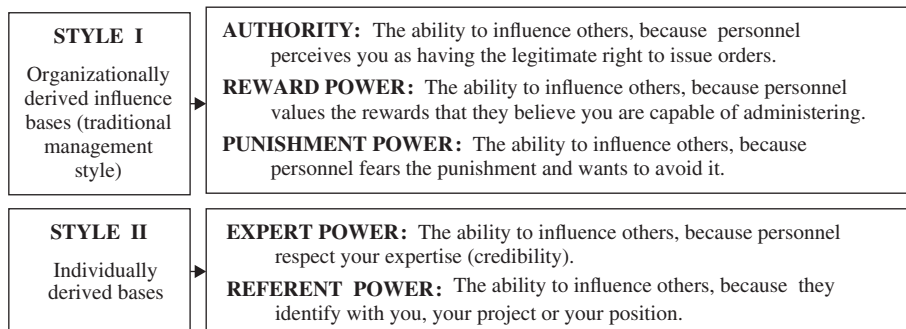


Figure 4 Common bases of managerial influence.

respect, credibility, and competence, it is strongly influenced by the manager's ability to foster a work environment where the team feels comfortable, accomplishes results, and receives recognition, and members have their professional and personal needs met.³¹ This includes images of managerial expertise, friendship, work challenge, promotional opportunities, fund allocations, charisma, personal favors, project goal identification, recognition, and visibility of the work and its importance.

Salary and other financial rewards play a very special role in the managerial power spectrum. It is often over or undervalued as a motivator. While overvaluation is costly and might set false expectations, ignoring or underplaying the importance of financial rewards can be detrimental to motivation and the morale of the people. It is quite common for engineering and technology managers to argue that salary and other financial rewards fulfill only lower level needs, while recognition, pride, and accomplishments are the true motivators of professional people. This is generally true. It is also supported by the writings in this chapter. However, the above argument holds only if personnel perceive the compensation as fair and adequate. Otherwise salary becomes a barrier to effective teamwork, a handicap for attracting and holding quality people, and a source of subtle conflict. To illustrate, a person who is motivated to make an extra effort might indeed enjoy the praise and recognition that comes with the well-done job. The person may further conclude that the job was important to the company and he or she is making significant contributions to the enterprise and could expect a better than average raise or be in line for a promotion. Now, suppose a subsequent salary review results in a very small, less than expected increase; the employee would question the sincerity and value of any praise, recognition, or other intrinsic rewards received in the past and anticipated for the future. The employee might also feel confused, frustrated, angry, or manipulated. Obviously, this is *not* a situation that leads to long-term motivation, sustained personal drive, commitment, and high morale. In summary, salary and other financial rewards are very important bases of managerial power. They must be used judiciously, consistent with the employee's output, efforts, and contributions, and be fair and equitable across the organization. Effectively employee communication, explaining the rationale for any financial award, supported with well-earned recognition of the accomplishments, is a good starting point for optimizing the motivational benefits of financial rewards.

Taken together, the sources of motivation involve an intricate set of variables which are grounded to a large degree in the work environment, team characteristics, and leadership style. Most importantly, managers must pay attention to the human side. To earn the trust, respect, and credibility of team members and support personnel, engineering managers must foster a work environment where people find the work interesting and challenging, leading to recognition and professional growth. Such a professionally stimulating environment seems to lower anxieties over managerial controls and conflict; it promotes cross-functional communication and enhances organizational awareness of the surrounding business environment, favorably affecting collaboration and the ability to unify the team behind the agreed-on objectives and organizational leadership.

5 MANAGING AND LEADING ENGINEERING TEAMS

Virtually all managers recognize the importance of effective teamwork critical to success in today's global hypercompetitive business environment.^{3,46-49} However, building and managing a work group as a fully integrated, unified team is a daunting task, especially for complex, technologically advanced, and geographically dispersed work groups which are common in engineering. This has been recognized by researchers and practitioners for a long time. In fact, the basic concepts go back to ancient times. Writings by Sun Tzu articulated specific approaches to teamwork and collaboration already 2500 years ago.⁵⁰ The first formal concepts evolved

with the *human relations movement* that followed Roethlingsberger and Dickinson's⁵¹ classic Hawthorne studies. Visionaries such as McGregor⁵² spelled out the criteria for effective group work (Theory Y), while Likert⁵³ called his highest form of management the participative group, or System 4 Management. Closer to the end of the last millennium, Dyer,⁵⁴ Tichy and Ulrich,⁵⁵ Walton,⁵⁶ Dumaine,⁵⁷ and Oderwald⁵⁸ have further broadened the understanding of team-based work processes.

Fast forward into today's more complex, multinational and technologically interactive engineering environment, the traditional work group reemerged as the *project team*, which can be defined as a collection of individuals selected for their specific skill sets and qualities. Often the group members have different needs, backgrounds, and experiences that must be skillfully focused and managed to transform the work group into an integrated, unified team.

In this transformation, referred to as *teambuilding*, the goals and energies of individual contributors merge and focus on specific objectives and desired results that characterize a high-performance team as summarized in Fig. 5. Building such a team requires sophisticated managerial skills. In today's complex business environment, many engineering teams are geographically dispersed or of multinational nature.^{6,8,12,28,59,60} This requires sophisticated communication, collaboration, and integration of information and activities among people from different organizations with different cultures, values, and languages. It also requires the ability to deal with uncertainties and risks caused by technological, economic, political, social, and regulatory factors, often across international borders. These concerns are also reflected in the large number of professional and executive education programs that have emerged in recent years to deal with these issues. Indeed, managing engineering operations is highly complex and difficult. From the senior management side, guidelines and unified direction toward project objectives, technology transfer, and project integration must be "synthesized and orchestrated" centrally and translated across borders into the cultures of the local operations.^{4,61} Then, linkages among individual work components need to be developed and effectively "managed" across geographic areas and organizational cultures. Thus, engineering teams not only need to be integrated across the miles but also need to be unified among different business processes, management styles, operational support systems, and organizational cultures.⁶²⁻⁶⁵

Yet, in spite of all these challenges, many enterprises competing in complex and technology-based ventures produce great results, even under extremely tight time and resource constraints. What lessons can we learn? Field studies of *technology-intensive environments*



Figure 5 Characteristics of high-performing teams.

point to two specific sets of variables: (1) *team leadership* and (2) *organizational ambiance*, strongly associated with team performance.³¹

5.1 Building High-Performance Teams

Part of the managerial challenge stems from the fact that work groups do not come as a fully developed team. Rather, project teams, in engineering or elsewhere, are being assembled based on their required skill sets. These people come from different organizations with different cultures, values, attitudes, personal preferences, and perceptions and work ethics. No matter how good and relevant their skill sets, these work groups need to be unified into a team that focuses its energy and effort on the mission objectives and desired results. This creates tough challenges, especially in technology-based organizations, where much of the work is team based and often distributed across wide geographic areas and teams include contractors, partner companies, regulators, and complex customer relations.

Because of these dynamics, power and responsibility are shifting from managers to project team members who take higher levels of responsibility, authority, and control within the work process and for specific results. That is, teams can rarely be managed “top down,” but become *self-directed*, gradually replacing the more traditional, hierarchically structured organization. These emerging team processes rely strongly on group interaction, resource and power sharing, group decision making, accountability, self-direction, and control. Organizing, building, and managing such self-directed teams require a keen understanding of the organization and its processes.⁶⁶ In addition, managers must realize the organizational dynamics involved during the various phases of the team development process. A four-stage model, originally developed by Hersey and Blanchard⁶⁷ and graphically shown in Fig. 6, can be useful as a framework to analyze and develop the work group toward a fully integrated high-performing team. The four stages are labeled (1) *team formation*, (2) *team start-up*, (3) *partial integration*, and (4) *full integration*. These stages are also known as *forming*, *storming*, *norming*, and *performing*, hence colloquially describing the team behavior at each one of the stages.

Applying the Four-Stage Team Development Model. Since no work group comes fully integrated and unified in its values and skill sets, team leaders need to nurture and build these groups from their *formation* through *start-up* to their *fully integrated phase iv*. To be effective, team

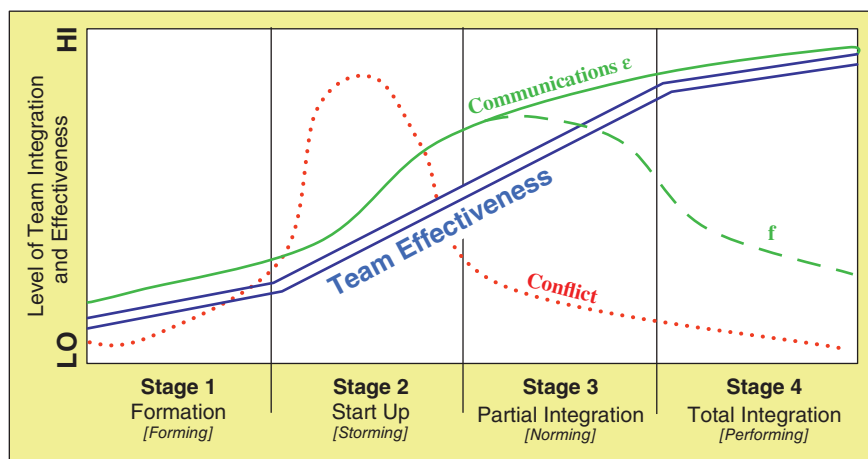


Figure 6 Four-stage development model.

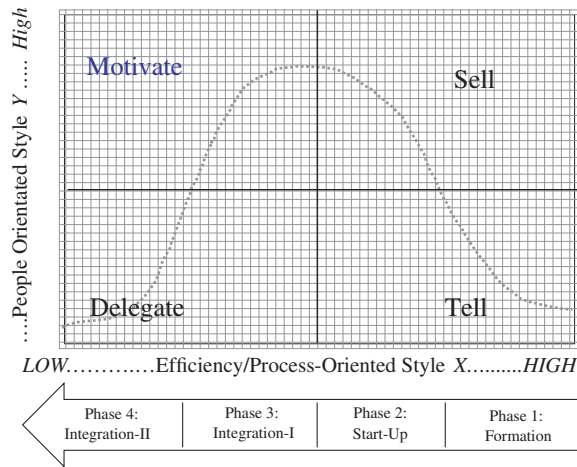


Figure 7 Team leadership at different phases of team development.

leaders must not only recognize the professional interests, needs, challenges, and anxieties of their team members as the group goes through the various stages of development but also adapt their managerial style to the specific situation that exists in each of the four stages.⁶⁸ That is, team leaders must recognize what works best at each stage and what is most conducive to the team development process. Many of the problems that occur during the formation of the new project team or during its life cycle are normal and often predictable. However, they present barriers to effective team performance. The problems must be quickly identified and dealt with. Early stages, such as the *team formation* and *start-up*, usually require a predominately directive style of team leadership. Providing clear guidelines on the project mission, its objectives, and its requirements and creating the necessary infrastructure and logistics support for the project team are critically important in helping the team to pass through the first two stages of their development quickly. During the third stage, *partial integration*, or *norming*, the team still needs a considerable amount of guidance and administrative support as well as support in dealing with the inevitable human issues of conflict, power and politics, credibility, trust, respect, and the whole spectrum of professional career and development. This is the stage where a combination of *directive and participative leadership* will produce most favorable results. Finally, a team that reached the fully integrated stage, by definition, becomes “self-directed.” That is, such a fully integrated, unified team can work effectively with a minimum degree of external supervision, administrative support, and managerial intervention. However, it requires highly sophisticated team leadership to maintain and sustain this delicate state of team effectiveness and focus. The situational leadership dynamics over the four stages of team building is graphically shown in Fig. 7.

6 RECOMMENDATIONS FOR EFFECTIVE ENGINEERING MANAGEMENT

The nature of engineering management, with its multidisciplinary technology requirements, ambiguous authority definition, and complex operating environment, requires experience and sophistication for effective role performance. A number of suggestions may be helpful to increase the managerial effectiveness.

Understand Motivational Needs. Engineering and technology managers need to understand the interaction of organizational and behavioral elements in order to build an environment conducive to their personnel's motivational needs. Two conditions seem to be especially critical to high engineering performance: professional interest and work support. However, identifying and satisfying these needs across a complex diversified work group are challenging and require special techniques and skills. Conventional tools, such as focus groups, action teams, suggestion systems, open-door policies, and management-by-wandering-around, complemented with computer-aided tools, such as PeopleSoft and online surveys, can provide a useful framework for identifying and profiling the needs of various segments of the engineering workforce.

Adapt Leadership to the Situation. Because their environment is temporary and often untested, engineering and technology managers should seek a leadership style that allows them to adapt to the often conflicting demands that exist within their organizations, support departments, customers, and senior management. They must learn to "test" the expectations of others by observation and experimentation. Leading an engineering team can rarely be done "top-down" but requires a great deal of interactive team management skills and senior management support. Although difficult, managers must be able to alter their leadership style as demanded by the specific work situation and its people.

Define the Team Structure, Work Process, and Communication Channels. Management must provide the infrastructure needed and conducive to effective cross-functional teamwork and technology transfer. This includes properly defined interfaces, task responsibilities, reporting relations, communication channels, and work transfer protocols. Most of the tools for systematically describing the work process and team structure come from the conventional project management system: project charter, defining the mission and overall responsibilities of the project organization, including performance measures and key interfaces; project organization chart, defining the major reporting and authority relationships; responsibility matrix or task roster; project interface chart, such as the N-Squared chart; and job descriptions and well-defined phase-gate criteria. All of these tools can help project managers develop cross-functional linkages, facilitate interdisciplinary cooperation, establish alliances, and improve communications. Although these tools have been used by project managers for a long time, they need to be fine tuned and calibrated to the specific engineering situation and carefully integrated with the overall business process and culture of the enterprise.

Build a High-Performance Image. Project teams that have a clear sense of purpose and confidence in their mission perform better. This is true for any environment. A high-performance image stimulates the team's interest, pride of participation, and sense of ownership. It serves as a bridging mechanism, helping to unify the team across the miles and cultures. It also builds professional confidence and encourages team members to reach out to think outside the box and to resolve issues "locally" with a minimum of administrative support. Project leaders and senior managers can build a favorable project image by making the project visible and stressing its importance via media exposure, management involvement, and budgetary actions as well as by emphasizing its critical success factors and the professional opportunities and rewards associated with the project. These factors promote project ownership among team members and encourage each individual's desire to succeed. To build such a high-performance image, team members must have a clear sense of the significance of their contributions. A good understanding of the organization's goals and mission objectives by all stakeholders is a critical prerequisite for team members to clearly see their roles and value added to the organization. A favorable project image is a catalyst for unifying the team behind the project objectives and for building commitment toward desired results.

Accommodate Professional Interests, Build Enthusiasm and Excitement. Engineering managers should try to accommodate the professional interests and desires of their personnel when negotiating tasks and during the execution. This leads to employee ownership and commitment, resulting in increased involvement, better communication, lower conflict, stronger work effort, and higher levels of creativity. Equally important, factors that satisfy professional interests and needs have a strong effect on team unification and overall project performance. Interesting work is a perception in the eyes of the employee that can be enhanced by the manager's actions, such as discussed above under the recommendation to build a high-performance image. While the scope of the work group may be fixed, the manager has the flexibility of allocating task assignments among various members. Well-established practices, such as front-end involvement of team members during the project planning or proposal phase and one-on-one discussions, are effective tools for matching team member interests and project needs.

Build Technical Expertise. Technological skill sets are the primary assets and competitive tools of an engineering enterprise. Managers must promote continuing education of their personnel to maintain engineering excellence and stay abreast of the latest technological advances. Equally important, managers of technology themselves must maintain technical expertise in their fields. Such an understanding is crucial for managers to win the confidence of their team members, build credibility with the customer communities, participate in the search for solutions, and lead a unified engineering/technology effort.

Plan Ahead. Effective planning early in the life cycle of a new technology program is highly recommended. Planning is a pervasive activity that leads to personnel involvement, understanding, and commitment. It helps to unify the task team, provides visibility, and minimizes future dysfunctional conflict.

Develop Organizational Interfaces and Communication Channels. Overall success of an engineering team depends on effective communications and cross-functional integration. Each task team should clearly understand its work interfaces and transfer mechanism, including those with outside contractors and suppliers. In addition to conventional technologies, such as voice mail, e-mail, electronic bulletin boards and conferencing, workspace design, regular meetings, reviews, and information sessions, can facilitate the free flow of information, both horizontally and vertically. Moreover, team-based reward systems can help to facilitate cooperation with cross-functional partners. QFD concepts, n -square charting, and well-defined phase-gate criteria can be useful tools for developing cross-functional linkages and promoting interdisciplinary cooperation and alliances.

Unify Management Process. Successful project integration requires a unified managerial process. While this critically important condition is widely recognized by managers and researchers (i.e., Ref. 28), it is difficult to achieve, especially for complex multidisciplinary and geographic dispersed work that involves a mixture of different cultures, management styles, and work processes. The policies, procedures, and protocols that define the management process and communication tools for linking the technology transfer must have both (i) mutual acceptability by all stakeholders throughout the project organization and (ii) enough flexibility to adapt to each task group. The specific challenge is to create a management process that functions consistently and reliably as a unified command-and-control system for the total project without disturbing the "local" norms and cultures that have deep roots and long timelines, a concern that is especially strong for multinational programs. A better understanding of the local culture and value system and proper sensitivity to the needs of the people make up an important starting point. While this mutual understanding, trust, and

respect are best developed by bringing people together in one place, collocation of the whole team is often difficult and costly to accomplish. Yet, many of the geographically dispersed project teams have managed to bring together key personnel from interfacing organizations, at least for short periods of time. Focus groups, organizational studies, interface developments, internal and external consultants, process action teams, professional training and teambuilding sessions, all are powerful tools for unifying and optimizing the work flow and for managing the process. These tools can be considerably enhanced through face-to-face contact, ideally at the beginning of the project life cycle, when both the people and the organizational processes are most flexible. Once mutual understanding, trust, and respect have been established, virtual communications using modern information technology might work quite well for fine tuning the implementation and maintenance of the work process and its management. However, the process of fine tuning and maintaining an effective multidisciplinary work process is an ongoing effort that requires extensive cross-organizational collaboration, continuous management involvement, and strong resource commitment.

Communicate Organizational Goals and Objectives. Management must communicate and update the organizational goals and project objectives. The relationship and contribution of individual work to the overall engineering project, the business plans, and the importance to the organizational mission must be clear to all team personnel. Senior management can help in unifying the team behind the project objectives by developing a “priority image” through their personal involvement and visible support.

Create Proper Reward Systems. Personnel evaluation and reward systems should be designed to reflect the desired power equilibrium and authority/responsibility sharing needed for the engineering organization to function effectively. Creating a system and its metrics for reliably assessing performance in engineering environment is a great challenge. However, models such as the integrated performance index provide a starting point for customization.⁶⁹

Build Commitment and Promote Self-Direction. Much of engineering management and the control of the work rely on commitment, motivation, and individual desire to succeed. However, building an environment conducive to self-direction and commitment requires more than simply empowering people. It requires creating an organizational environment where people feel confident and enthusiastic about their work. The proper organizational structure, expertise and high-performance image, and many of the conditions discussed in this section provide the basis for such a favorable environment. Conflict over technical, administrative, or personal issues, lack of interest, anxieties, and fear of the unknown are often reasons for low or unsustainable commitment,⁷⁰ which should be carefully monitored and rectified before they affect team performance.

Manage Conflict and Problems. Engineering activities are highly disruptive to the enterprise and conflict is inevitable. Managers should focus their efforts on early problem identification avoidance. That is, managers and team leaders, through experience, should recognize potential problems and conflicts at their onset and deal with them before they become big and consume a large amount of time and effort.

Conduct Team Building Sessions. A mixture of *focus group* sessions, *brainstorming*, *experience exchanges*, and *social gatherings* can be powerful tools for developing the work group into an effective, fully integrated, and unified project team.⁷¹ Intensive team building efforts may be needed, especially during the formation stage of a new project team, but should be conducted throughout a project life cycle.

Foster a Culture of Continuous Support and Improvement. Successful engineering management focuses on people behavior and their roles within the organization. High-performance engineering organizations have cultures and support systems that demand broad participation in their organization developments. It is important to establish support systems—such as discussion groups, action teams, and suggestion systems—to capture and leverage the lessons learned and to identify problems as part of a continuous improvement process. Tools such as the *Project Maturity Model*^{72,73} and the *Six Sigma Management Process*^{74,75} can provide the framework and tool set for analyzing and fine tuning the team development and its management process.

Provide Effective Direction and Leadership. Engineering managers and team leaders can influence the attitude and commitment of their people by their own actions and leadership style. Concern for team members, the ability to integrate personal and organizational goals, and the manager's enthusiasm for the project can foster a climate of high motivation, work involvement, open communications, creativity, commitment, and ultimately high engineering/technology performance.

7 CONCLUDING REMARKS

Managerial leadership has significant impact on the work environment, affecting engineering personnel and performance. Factors that satisfy personal and professional needs—such as *professional interest, pride and satisfaction with the work, professional work challenge, accomplishments and recognition*—seem to have the most favorable impact on individual and team performance, reducing communication barriers while enhancing collaboration and the desire to succeed. Other important influences include effective communications among team members and support units across organizational lines, good team spirit, mutual trust and respect, low interpersonal conflict, and opportunities for career development and job security. These conditions serve as a bridging mechanism between personal and organizational goals and are helpful in building a unified engineering team capable of producing integrated results in support of the organization's mission. Effective engineering managers are social architects who understand the interaction of organizational and behavioral variables who can foster a climate of active participation, accountability, and result orientation throughout the enterprise and its partners. This requires a keen in-depth understanding of the organizational dynamics and its cultures plus sophisticated technology management and leadership skills. It further requires the ability to engage top management, ensuring organizational visibility, resource availability, and overall support for the engineering work across the enterprise.

REFERENCES

1. D. Anconda and H. Bresman, *X-teams: How to Build Teams That Lead, Innovate and Succeed*, Harvard Business School Publishing Company, Boston, MA, 2007.
2. I. Kruglianskas and H. Thamhain, "Managing Technology-Based Projects in Multinational Environments," *IEEE Trans. Eng. Manag.*, **47**(1), 55–64, 2000.
3. R. Nellore and R. Balachandra, "Factors Influencing Success in Integrated Product Development (IPD) Projects," *IEEE Trans. Eng. Manag.*, **48**(2), 164–173, 2001.
4. A. Nurick and H. J. Thamhain, "Team Leadership in Global Project Environments," in D. I. Cleland (Ed.), *Global Project Management Handbook*, McGraw-Hill, New York, 2006, Chapter 38.
5. B. Schmid and J. Adams, "Motivation in Project Management: A Project Manager's Perspective," *Project Manag. J.*, **39**(2), 60–71, 2008.
6. A. Shenhar, "What Great Projects Have in Common," *MIT Sloan Manag. Rev.*, **52**(3, Spring), 19–21, 2011.

7. S. Sidle, "Building a Committed Workforce: Does What Employers Want Depend on Culture?" *Acad. Manag. Perspect.*, **23**(1), 79–80, 2009.
8. H. Thamhain, "Critical Success Factors for Managing Technology-Intensive Teams the Global Enterprise," *Eng. Manag. J.*, **23**(3, September), 30–36, 2011.
9. D. Armstrong, "Building Teams Across Borders," *Exec. Excell.*, **17**(3), 10–11, 2000.
10. H. Barkema, J. Baum, and E. Mannix "Management Challenges in a New Time," *Acad. Manag. J.*, **45**(5), 916–930, 2002.
11. R. Barner, "The New Millennium Workplace," *Eng. Manag. Rev. IEEE*, **25**(3, Fall), 114–119, 1997.
12. A. Bhatnager, "Great Teams," *Acad. Manag. Exec.*, **13**(3 (August)), 50–63, 1999.
13. P. Dillon, "A Global Challenge," *Forbes Magazine*, **168**(September 10), 73+, 2001.
14. C. Gray and E. Larson, *Project Management*, Irwin McGraw-Hill, New York, 2000.
15. H. Thamhain, "Criteria for Effective Leadership in Technology-Oriented Project Teams," in D. Slevin, D. Cleland, and J. Pinto (Eds.), *The Frontiers of Project Management Research*, Project Management Institute, Newton Square, PA, 2002, pp. 259–270.
16. R. Keller, "Cross-Functional Project Groups in Research and New Product Development," *Acad. Manag. J.*, **44**(3), 547–556, 2001.
17. S. Manning, S. Massini, and A. Lewin, "A Dynamic Perspective on Next-Generation Offshoring: The Global Sourcing of Science and Engineering Talents," *Acad. Manag. Perspect.*, **22**(3), 35–54, 2008.
18. F. Newell, and M Rogers, *loyalty.com: Relationship Management in the Era of Internet Marketing*, McGraw-Hill, New York, 2002.
19. H. Thamhain, "Leadership Lessons from Managing Technology-Intensive Teams," *Int. J. Innovation Technol. Manag.*, **6**(2), 117–133, 2009.
20. Drucker, P. 2006. *Innovation and Entrepreneurship*. New York: Harper Collins.
21. D. Anconda, T. Malone, W. Orlikowski, and P. Senge "In Praise of the Incomplete Leader," *Res. Technol. Manag.*, **19**(3, May–June), 92–100, 2007.
22. M. Hilton, "Skills for Work in the 21st Century," *Acad. Manag. Perspect.*, **22**(4), 63–78, 2008.
23. M. Sawhney, "Don't Just Relate—Collaborate," *MIT Sloan Manag. Rev.*, **43**(3), 96–107, 2002.
24. M. Sawhney and E. Prandelli, "Communities of Creation: Managing Distributed Innovation in Turbulent Markets," *Calif. Manag. Rev.*, **42**(4), 45–69, 2000.
25. C. Tomkovich and C. Miller, "Riding the Wind: Managing New Product Development in the Age of Change," *Product Innovation Manag.*, **17**(6, November), 413–423, 2000.
26. M. Debruyne, R., Moenaert, A. Griffin, and S. Hart, *J. Product Innovation Manag.*, **19**(2, March), 159–169, 2002.
27. J. Hackman, *Leading Teams: Setting the Stage for Great Performance—The 5 Keys to Successful Teams*, Harvard Business School Press, Boston, 2002.
28. J. Hackman, "The Five Dysfunctions of a Team: A Leadership Fable," *Acad. Manag. Perspect.*, **20** 122–125, 2006.
29. J. Solomond, "International High Technology Cooperation: Lessons Learned," *IEEE Trans. Eng. Manag.*, **43**(1, February), 69–78, 1996.
30. H. Thamhain, "Managing Innovative R&D Teams," *R&D Manag.*, **33**(3, June), 297–312, 2003.
31. H. J. Thamhain, "Team Leadership Effectiveness in Technology-Based Project Environments," *IEEE Eng. Manag. Rev.*, **36**(1), 165–180, 2008.
32. N. Arranz and J. de Arroyabe, "Joint R&D Projects as Complex Systems," *IEEE Trans. Eng. Manag.*, **55**(4), 552–566, 2008.
33. D. Anconda, T. Malone, W. Orlikowski, and P. Senge "It's Time to End the Myth of the Incomplete Leader," *Harvard Bus. Rev.*, **85**, 92–100, 2007.
34. D. Cohen, "Interview with Alexander Laufer," *Acad. Sharing Knowledge, ASK*, Issue 35 (Summer 2009), 23–28, 2009.
35. M. Hoegl, H. Ernst, and L. Proserpio, "How Teamwork Matters More as Team Member Dispersion Increases," *J. Product Innovat. Manag.*, **24**(2), 156–165, 2007.
36. H. S. Wade (Ed.), Special Issue on "Leading Small Groups," *IEEE Trans. Eng. Manag.*, **37**(3), 3–86, 2009.

37. A. Shenhar "Strategic Project Leadership: Toward a Strategic Approach to Project Management." *R&D Manag.*, **34**(November), 569–578, 2004.
38. M. Dayan, S. Elbanna, and A. Di Benedetto, "Antecedents and Consequences of Political Behavior in New Product Development Teams," *IEEE Trans. Eng. Manag.*, **59**(3), 470–482, 2012.
39. P. Patanakul and A. Shenhar, "What Is Really Project Strategy: The Fundamental Building Block in Strategic Project Management." *Project Manag. J.*, **43**(1, February), 4–20, 2012.
40. H. J. Thamhain, "Leading Technology Teams," *Project Manag. J.*, **35**(4), 35–47. (2004).
41. H. J. Thamhain, "Team Leadership Effectiveness in Technology-Based Project Environments," *IEEE Trans. Eng. Manag.*, **33**(2), 11–25, 2005.
42. D. Cleland and L. Ireland, *Project Management: Strategic Design and Implementation*. McGraw-Hill, New York, 2007.
43. A.R. Jassawalla, R. Avan, and H.C. Sashittal, "Building Collaborate Cross-Functional New Product Teams," *Acad. Manag. Exec.*, **13**(3), 50–63, 1999.
44. J. Polzer, C. Crisp, S. Jarvenpaa, and J. Kim, "Extending the Faultline Model to Geographically Dispersed Teams," *Acad. Manag. J.*, **49**(4), 679–692, 2006.
45. J. R., French, Jr., and B. Raven, "The Basis of Social Power," in D. Cartwright (Ed.), *Studies in Social Power*, Research Center for Group Dynamics, Ann Arbor, MI, 1959, pp. 150–165.
46. C. Ferrante, S. Green, and W. Forster, "Getting More Out of Team Projects: Incentivizing Leadership to Enhance Performance," *J. Manag. Ed.*, **30**(6), 788–798, 2006.
47. B. Groysberg and R. Abrahams, "Lift Outs: How to Acquire a High-Functioning Team," *Harvard Bus. Rev.*, **84**(12), 133–143, 2006.
48. M. Hoegl and K. P. Parboteeah, "Creativity in Innovative Projects: How Teamwork Matters," *J. Eng. Technol. Manag.*, **24**, 148–166, 2007.
49. D. Shim and M. Lee, "Upward Influence Styles of R&D Project Leaders," *IEEE Trans. Eng. Manag.*, **48**(4), 394–413, 2001.
50. T. Hanzhang, *Sun Tzu's art of war: the modern Chinese interpretation*, Sterling Publishing Company, New York, 2007.
51. F. Roethlisberger and W. Dickerson, *Management and the Worker*, Harvard University Press, Cambridge, MA, 1939.
52. D. McGregor, *The Human Side of Enterprise*, McGraw-Hill, New York, 1960.
53. R. Likert, *The Human Organization*, McGraw Hill, New York, 1967.
54. W. G. Dyer, *Team Building: Issues and Alternatives*, Addison-Wesley, Reading, MA, 1977.
55. N. M. Tichy and D. O. Ulrich, "The Leadership Challenge—A Call for the Transformational Leader," *Sloan Manag. Rev.*, **26**, 59–68, 1984.
56. R. Walton, "From Control to Commitment in the Workplace. *Human Resource Management: Critical Perspectives on Business and Management*, **1**(15), 1999.
57. B. Dumaine, "The Bureaucracy Buster," *Fortune*, June 17, 1991.
58. S. Oderwald, "Global Work Teams." *Training and Development* **5**(2), 1996.
59. K. Brockhoff and B. Schmaul, "Organization, Autonomy, and Success of Internationally Dispersed R&D Facilities," *IEEE Trans. Eng. Manag.*, **43**(1, February), 33–40, 1996.
60. S. Ohba, "Critical Issues Related to International R&D Programs," *IEEE Trans. Eng. Manag.*, **43**(1, February), 78–87, 1996.
61. H. Thamhain, "Managing Globally Dispersed R&D Teams." *Int. J. Inform. Technol. Manag. (IJITM)*, **8**(1), 107–126, 2009.
62. H. Bahrami, "The Emerging Flexible Organization: Perspectives from Silicon Valley," *California Management Review*, **34**(4), 33-52, 1992.
63. A. De Maio, R. Verganti and M. Corso, "A Multi-Project Management Framework for New Product Development," *European Journal of Operational Management*, **78**(2), 178-191, 1994.
64. J. Deschamps and R. Nayak, "Implementing World-Class Process," Chapter 5 in *Product Juggernauts* (Deschamps, ed.), Cambridge: Harvard Press, 1995.
65. M. Gibbert and M Hoegl, "In Praise of Dissimilarity," *Sloan Manag. Rev.*, **52**(4), 20–22, 2011.
66. G. Cutler and R. Smith, "Mike Leads His First Virtual Team," *Res. Technol. Manag.*, **50**(1), 66–69, 2007.

67. P. Hersey and K. Blanchard, *Management of Organizational Behavior*, Prentice Hall, Englewood Cliffs, NJ, 1996.
68. H. Thamhain and D. Wilemon, "Building High Performing Engineering Project Teams," in R. Katz (Ed.), *The Human Side of Managing Technological Innovation*, Chapter 12, Oxford University Press, New York, 1996.
69. S. A. Pillai, A. Joshi, and K. S. Rao, "Performance Measurement of R&D Projects in a Multi-Project, Concurrent Engineering Environment," *Int. J. Project Manag.*, **20**, 165–177, 2002.
70. D. Stum, "Maslow Revisited: Building the Employee Commitment Pyramid," *Strategy and Leadership*, **29**(4), 4–9, 2001.
71. H. Thamhain and D. Wilemon, "Building Effective Teams in Complex Project Environments," *Technol. Manag.*, **5**(2, May), 203–212, 1999.
72. K. Crawford, *Project Management Maturity Model*, Taylor & Francis, Boca Raton, FL, 2007.
73. S. Fahrenkrog, F. Abrams, W. Haeck, and D. Whelbourne, "Project Management Institute's Organizational Project Management Maturity Model OPM3™," in *Proceedings of 2003 PMI North American Congress*, Baltimore, MD, Project Management Institute, Newtown, PA, 2003.
74. S. Neuendorf, *Six Sigma for Project Managers*, Management Concepts, Vienna, VA, 2004.
75. N. Arranz and J. de Arroyabe, "Joint R&D Projects as Complex Systems," *IEEE Trans. Eng. Manag.*, **55**(4), 552–566, 2008.
76. P. Drucker, *Innovation and Entrepreneurship*, Harper Collins, New York, 2006.
77. K. Fisher, *Leading Self-Directed Work Teams*, McGraw-Hill, New York, 1993.
78. M. Hoegl and K. P. Parboteeah, "Team Goal Commitment in Innovative Projects," *Int. Natl. J. Innovat. Manag.*, **10**(3), 299–324, 2006.
79. D. Milosevic (Ed.), *Project Manager's Tool Box*, Wiley, New York, 2003.
80. S. Oderwald, "Global Work Teams," *Training Devel.*, **5**(2, February), 1996.
81. A. Schulze and M. Hoegl, "Knowledge Creation in New Product Development Projects," *J. Manag.*, **32**(2), 210–236, 2006.
82. P. Senge and G. Carstedt, "Innovating Our Way to the Next Industrial Revolution," *Sloan Manag. Rev.*, **42**(2), 24–38, 2001.
83. P. Senge, *The Fifth Discipline: The Art and Practice of the Learning Organization*, Doubleday/Currency, New York, 1994.
84. B. Sharma, "R&D Strategy and Australian Manufacturing Industry: An Empirical Investigation of Emphasis and Effectiveness," *Technovation*, **23**(12, December), 929–937, 2003.
85. A. Shenhar, D. Dvir, D. Milosevic, and H. Thamhain, *Linking Project Management to Business Strategy*, Project Management Institute (PMI) Press, Newtown, PA, 2007.
86. J. B. Schmidt and R. J. Calantone, "Escalation of Commitment during New Product Development," *J. Acad. Marketing Sci.*, **30**(1), 103–118, 2002.
87. H. Thamhain, "The Future of Project Team Leadership," B. Bidanda and D. Cleland, Eds.). *Project Management Circa 2025* (professional reference book), PMI Press, Philadelphia, PA, 2009, Chapter 11.
88. L., Valikangas, M. Hoegl, and M. Gibbert, "Why Learning from Failure Isn't Easy (and What to Do About It): Innovation Trauma at Sun Microsystems," *Eur. Manag. J.*, **27**(4), 225–233, 2009.
89. V. Verma, "Conflict Management," in J. Pinto (Ed.), *Project Management Handbook*, Jossey Bass, San Francisco, 1998, pp. 353–376.
90. N. Whitten, *Managing Software Development Projects*, 2nd ed., Wiley, New York, 1995.
91. C. Zafft, S. Adams, and G. Matkin, "Measuring Leadership in Self-Managed Teams Using the Competitive Value Framework," *J. Eng. Ed.*, **98**(3), 273–282, 2009.
92. R. Zaroni, and J. Audy, "Project Management Model for Physically Distributed Software Development Environment," *Eng. Manag. J.*, **16**, 1, 2004.

CHAPTER 19

ENGINEERING ECONOMY

Kate D. Abel
Stevens Institute of Technology
Hoboken, New Jersey

1 INTRODUCTION - WHAT IS ENGINEERING ECONOMICS?	581	8.4 Internal Rate of Return	595
2 IMPORTANCE OF ETHICS IN ENGINEERING ECONOMICS	582	8.5 Benefit–Cost Analysis	597
3 TYPES OF ENGINEERING ECONOMIC DECISIONS	583	9 CAPITAL RECOVERY, CAPITALIZED COST, AND REPLACEMENT STUDIES	598
4 CASH FLOWS AND TIME VALUE OF MONEY	583	9.1 Capital Recovery	598
5 EQUIVALENCE	584	9.2 Capitalized Cost	599
5.1 Cash Flow Diagrams, Interest Rates, and Setting Up the Problem	584	9.3 Replacement Studies	599
6 SINGLE SUM AND UNIFORM, GRADIENT, AND GEOMETRIC SERIES	588	10 ADDITIONAL ANALYSES AND CONSIDERATIONS IN THE SELECTION PROCESS	599
7 COMPARING ALTERNATIVES: DEFINING ALTERNATIVES	591	10.1 Depreciation	599
8 COMPARING ALTERNATIVES THROUGH THE FIGURES OF MERIT	592	10.2 Inflation	600
8.1 Present Value	592	10.3 After Tax Analysis	601
8.2 Future Value	592	10.4 Sensitivity Analysis	601
8.3 Annual Worth	592	10.5 Break-Even Analysis	602
		10.6 Risk Analysis	602
		11 CONCLUSION	603
		REFERENCES	603

1 INTRODUCTION - WHAT IS ENGINEERING ECONOMICS?

Imagine you were on a team tasked by the New York City Metro Transit Authority to design the next way to allow people to commute across the Hudson River between New York City and New Jersey. What would be your response? Would your solution be a tunnel? A bridge? A train tunnel? A car tunnel? A bridge allowing both trains and cars across it? What about a ferry system? Would you send people over by catapult? Or helicopter? Or airplane? Obviously, some of these choices are technologically more sound than others. But once you came down to whether it should be a tunnel, a bridge or a ferry system, how would you decide between the final models?

What would be the method used to choose between the very different technologically possible solutions?

Engineering Economics provides the tools needed to compare these varied technologies on a level playing field. No matter what technologies or technological solutions are being compared—whether they be the same or very different—often the solution chosen is the one of least cost. Engineering Economics, therefore, provides a rational selection process to determine how to allocate capital.

Consider instead alternative fuel vehicles. Engineers designing these vehicles must consider acceleration and engine power versus fuel efficiency; vehicle safety and maintenance versus material choice; distance between refuels versus tank size, etc.; in addition to all sorts of technical decisions, as well as, ethical issues. All these design decisions influence manufacturing costs and therefore the final sales price. And therefore it can be seen that the entire supply, production, and service chains would be affected by the design choices too. Ultimately, the entire profitability, or lack thereof, of the new alternative fuel vehicle is affected by the design choices of the engineer. How can the engineer assist in helping his design be a profitable one? Consider Engineering Economics during the design process. Consider the financial impact of a decision while making choices between alternative ways to solve the technical problem. That is Engineering Economics.

What other kinds of questions can Engineering Economics answer in the business realm?

- Which technical projects should be considered of higher priority?
- Has the industrial engineer analyzed which of the proposed improvement projects on the production line should be funded with the money allocated in the budget?
- Which technical projects are worth pursuing and which are not?
- Has an analysis been done demonstrating that the purchase of a new machine will indeed cost the company less over time than keeping the current older version?

Engineering Economics can be used to answer personal questions as well. For example, is it better to finance your car through the dealership? Or is it less expensive in the long run to take the rebate and borrow money from your bank? Another example would be how much money should you put in your IRA each month now, so when you retire you have \$3 million? From car loans and mortgages to long-term financial goals and simply understanding finances, Engineering Economic analyses can assist one in making the choice that has the best financial impact on you, your business, and your family.

2 IMPORTANCE OF ETHICS IN ENGINEERING ECONOMICS

To say that engineering economics assists in making the best decision among alternatives from a financial perspective, also begs one to additionally discuss ethics. For what if the safest solution is also the most expensive? And the least expensive option toxic, but chosen anyway due to cost savings. This could be what happened to Mattel in 2007 when some toys made in China were coated with lead paint. Was the use of lead paint an economic decision made by Chinese subcontractors to increase their bottom line?

Engineers must be mindful of the ethical dimensions of their technical and design decisions. Usually the ethical choice is relatively clear. But with conflicting moral imperatives, sometimes the “right” decision can create an ethical dilemma. To assist in this, many engineering professions have professional codes that serve as a reference for both new engineers and the legal system. The National Society of Professional Engineers (NSPE) has a code of ethics. One of the gravest responsibilities of any engineer and the first NSPE fundamental canon of ethical behavior is to “hold paramount the safety, health and welfare of the public.” But a cost versus profit mentality muddies up the waters in decision making. For example, should a product

go through an additional month of testing before being released? In determining standards for pollutants, is 1 part per million acceptable, or should more stringent levels be set? Should a company use the best subcomponents in the manufacture and assembly of its products? Or, by using lower quality and thus lower cost parts, does it improve the firm's profitability while reducing the usable life of the product at the same time? Often engineers try to act ethically, however, designs change, data is incorrect, or a choice is made between profits and ethics. Companies are driven by profit. Often it is left to the engineer to identify at what point safety is being compromised.

3 TYPES OF ENGINEERING ECONOMIC DECISIONS

There are two main types of Engineering Economic decisions. The first is a capital expenditure decision. Money exists, normally a finite amount from a budget allocation. This money must be spent in the most optimal way; one wants to maximize the benefits derived from the capital. Using Engineering Economic techniques, one can determine the best way to allocate the capital across projects or design alternatives: whether equipment should be repaired or replaced, the best equipment choice among alternative solutions, are iterative improvements in design worth the cost in manufacturing?

The second kind of Engineering Economic decision involves cost reduction. In this case, there are normally no profits, but a change in state, or an optimal choice between alternatives to reduce costs. These types of problems are referred to as "least cost" problems since one is looking to find the least costly solution among alternatives. An example of this type of problem might be the installation of adequate lighting in a new office space. New lighting must be purchased for employees to see and work, but the lighting itself will provide no direct profits for the company. However, the choice of installation between halogen, florescent or compact florescent fixtures, etc. can impact costs over the life of the building. One type of fixture may have an expensive installation cost but low energy consumption costs over the life of the fixture. Another choice may cost little in installation, but utility bills may be very high over the life of the fixtures.

4 CASH FLOWS AND TIME VALUE OF MONEY

Money makes, or loses, value over time. Transactions, where money moves from one entity to another, are called cash flows. These cash flows can occur at any point in time or over the course of a time period, such as the case with a loan or mortgage. These transactions occur because the individual wants to modify her wealth—whether it is through putting the money in the bank, or taking out a mortgage, or investing in a new piece of equipment.

But what is the value of the money in these transactions? Due to interest rates and loss of purchasing power over time, the actual value of money changes. Take, for example, a loan taken out for a 30-year home mortgage at a specific interest rate. The borrower normally pays back the same amount in year 1 as he does in year 30. Although the dollar amounts are the same, the actual value of the money is very different. For example, your parents probably took out a mortgage for their home in the last decades of the twentieth century. At that point a \$1000 monthly payment was quite a hefty sum. In approximately 5–10 years, when your parents are in their final year of repayment, a mortgage payment of \$1000 a month will seem like pennies, compared to what the young couple next door is paying a month for a home of similar size that they just purchased. The reason for this is that over the life of your parent's 30-year mortgage, inflation has increased the price of homes, while the purchasing power of the \$1000 your parent's paid to the mortgage company each month has decreased. Thus the future value of the \$1000 is much less, 30 years later.

Therefore, not only is a monetary exchange needed to quantify the time value of money, but a time period and an interest rate must be known as well. As can be seen in the example above, \$1000 thirty years ago is not worth the same as \$1000 today. To equate the two would not be accurate, as it would not account for the “time value of money.” Thus any comparison between the two dollar amounts would need to be corrected through Engineering Economic techniques—equating sums of money from different periods of time based on the risk associated with the investment, normally expressed as an interest rate.

5 EQUIVALENCE

Consider two monetary transactions, if you are indifferent to choosing one or the other, then they are equivalent. Said another way, if you could either receive a quantity of money now or the assurance of some other sum of money in the future, and you are indifferent to choosing one over the other, the two sums are said to be equivalent. A lender–borrower scenario would be a typical example of this. Typically, when one borrows money, the borrower has to pay back more than that which was borrowed. This extra money could be thought of as “rent” for using the money, just as one might pay rent for using someone else’s home. To account for this in engineering economics, we use interest rates to indicate the value of this extra money over the period of the loan. Therefore, the amount the borrower borrows today and the amount that the lender receives in return over some future time period are “equivalent” in that these two amounts are related to each other by some interest rate and time period.

Under the concept of equivalence, all monetary sums must be moved to coincide with the same point in time. The time period can either be present or future, but not both. Therefore, either tomorrow’s dollars need to be converted to today’s worth or today’s dollars need to be converted to an equivalent future sum. This equivalence allows one to convert from any present value to any future value and vice versa. An example of this relationship is the lottery. The lottery will pay out either a lump sum today or multiple payouts over some specified time period in the future. These two payouts are equivalent at some interest rate. If we assumed that the lottery winner could only reinvest his winnings at the interest rate the lottery was paying out, then the lottery winner could be said to be indifferent to the two choices. (This example does not take into account specific exceptions such as income taxes, etc.)

5.1 Cash Flow Diagrams, Interest Rates, and Setting Up the Problem

Table 1 lists the five main elements involved in the time value of money: If one was just to look at the first four elements, equivalence dictates that the future sum F is an amount equivalent to P at some interest rate i , N periods away from the present. Cash flow diagrams are a graphical representation of these types of engineering economic problems. Similar to free-body diagrams, they are an important step in transforming a word problem to a form that can be analyzed using mathematical techniques. The cash flow diagrams graphically display the above elements,

Table 1 Five Main Elements in the Time Value of Money

	Eng Econ Notation	Notation Used in Excel
Interest rate per time period	i	Rate
Number of time periods in the planning horizon	N	Nper
Present sum of money (at time $N = 0$)	P	PV
Future sum of money (at some time $N = n$)	F	FV
Annuity or uniform payment over time	A	Pmt

as well as others introduced later. The x axis designates the time periods N , while the y axis designates the amount of the monetary sums. Up represents a positive cash inflow, and down represents a cash outflow or expense. Each arrow represents a monetary sum changing hands. Although continuous cash flow analyses can be performed, more typical analyses assume the cash flows occur at the end of the time period. This designation is called the “end-of-year” convention and is the generally accepted format for cash flow diagrams. Please see Fig. 1a and 1b for examples of cash flow diagrams. See Fig. 1c for an example of drawing cash flow diagrams with spreadsheets.

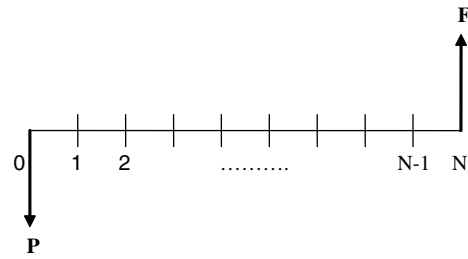
The typical steps to solve a cash flow analysis are straightforward. First, one should read the problem and identify the knowns and unknowns. Draw a cash flow diagram including all the components given in the problem, including those for which you are looking. Convert all knowns to the same units of time. For instance, if you are told that your mortgage is 6% per year for 30 years, but the payments you are searching for will be monthly, you need to convert everything to months. Next identify the formula or Excel function. The general equations used in engineering economics as well as the primary Excel functions are shown in Tables 1 and 2. These are commonly available in the literature¹ and on the internet.² And finally, solve the problem and make a recommendation using Engineering Economic techniques.

The formulas can also be written in factor notation as shown in the notation column of Table 2. The notation in the parentheses is available in tabular format such as Table 3 in the literature³ or on the Internet². Which interest rate to be used and which factor to be used will depend upon the interest rate given in the problem and the knowns and unknowns presented in the problem statement. Normally, each table is based on one interest rate or minimum attractive rate of return (MARR). Common mathematical operations would lead one to choose the appropriate factor, which would “cancel out” the known item on the right-hand side of the equation. Although this is not what actually happens, it allows one to easily choose the correct factor by which to multiply. In addition to determining the interest rate and factor, one would need to know the planning horizon, N , which runs down the first column of Table 3. Although many of the common interest rates have interest rate tables available in print or on the Internet, in the cases when they are not available one would need to use the formulas to calculate the solution.

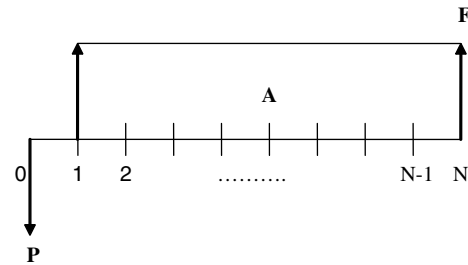
When considering interest, it should be noted that the interest rate is assumed to be compound interest (versus simple interest). Also, the interest rate must be the effective interest rate per period N . This is in line with what was said two paragraphs earlier that the time periods must align in all components used in the equation. Therefore, as the example above stipulated, one cannot use a yearly interest rate when the time period N is in months.

Finally, there are three different kinds of interest rates. The periodic rate ($i_m = r/m$) is the rate of interest within a specific compounding cycle, like within one quarter or one month. The nominal rate ($r = m \times i_m$) is the periodic rate scaled to a yearly basis (m = number of interest periods per year). So if the periodic rate was 2% per quarter, the nominal rate would be 8% per year ($r = 2\% \times 4$ quarters in a year). In other words the nominal rate is the rate per year calculated from straight extrapolation of the periodic rate. The effective rate is the equivalent of the periodic rate compounded for the number of compounding periods within the year. The effective rate will always be higher than the nominal rate since the effective rate takes compounding into account.

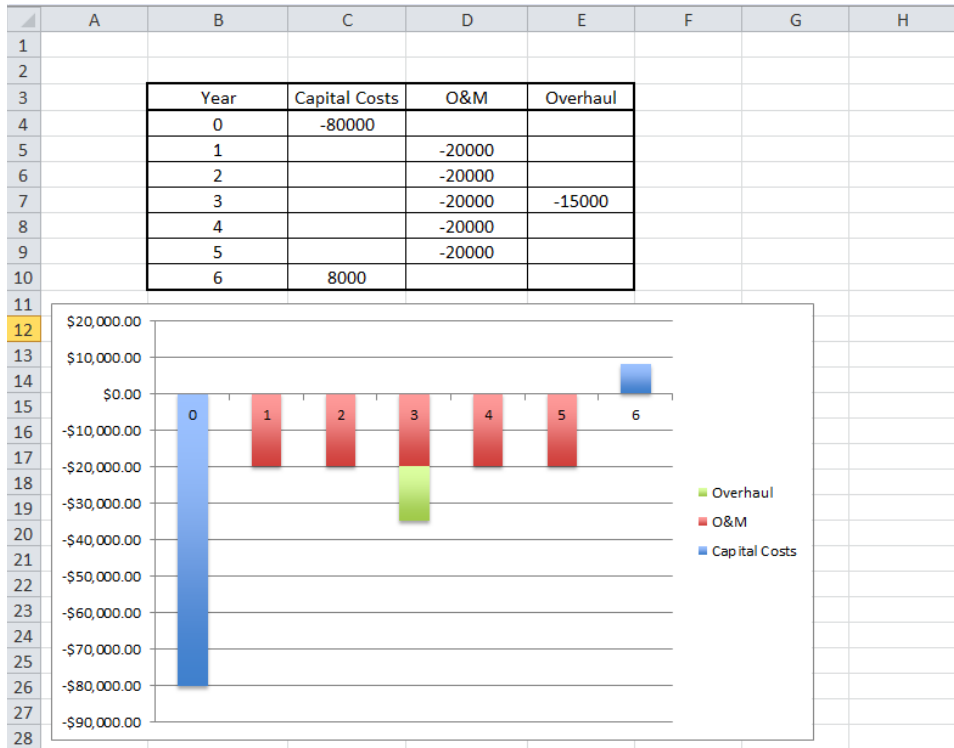
In many instances one can assume the rate given in a problem to be the effective rate if the compounding is annual. However, most real-world problems have interest compounding periods other than yearly. In fact, most interest is compounded daily, while the payments are made monthly. To use the formulas in Table 2, the payments must be in the same units of time as the interest. To convert the interest to various time units, additional notations must be introduced (k = number of payment periods per year and C = number of interest periods per payment



(a)



(b)



(c)

Figure 1 (a) Example of typical cash flow diagram—single sum (b) example of typical cash flow diagram—uniform series; and (c) example of drawing a cash flow diagram with spreadsheets.

Table 2 Discrete Cash Flow Formulas

	Type of Factor	Factor Symbol	Factor Notation	Formula
Present worth	Single payment	P/F	$(P/F, i, N)$	$\frac{1}{(1+i)^N}$
	Annuity (uniform series)	P/A	$(P/A, i, N)$	$\frac{(1+i)^N - 1}{i(1+i)^N}$
	Arithmetic gradient	P/G	$(P/G, i, N)$	$\frac{(1+i)^N - 1}{i^2(1+i)^N} - \frac{N}{i(1+i)^N}$
	Geometric gradient		$(P/A, i, g, N)$	$\frac{1 - [(1+g)/(1+i)]^N}{i-g}$
Future worth	Single payment	F/P	$(F/P, i, N)$	$(1+i)^N$
	Annuity (uniform series)	F/A	$(F/A, i, N)$	$\frac{(1+i)^N - 1}{(i)}$
	Arithmetic gradient	F/G	$(F/G, i, N)$	$\frac{(1+i)^N - 1}{i^2} - \frac{N}{i}$
Annuity	Capital recovery	A/P	$(A/P, i, N)$	$\frac{i(1+i)^N}{(1+i)^N - 1}$
	Sinking fund	A/F	$(A/F, i, N)$	$\frac{i}{(1+i)^N - 1}$
	Arithmetic gradient	A/G	$(A/G, i, N)$	$\frac{1}{i} - \frac{N}{(1+i)^N - 1}$

Table 3 Six Percent Interest Rate

	Compound Amount Factor	Present Worth Factor	Compound Amount Factor	Sinking-Fund Factor	Present Worth Factor	Capital Recovery Factor	Uniform Gradient Series Factor
N	$(F/P, i, N)$	$(P/F, i, N)$	$(F/A, i, N)$	$(A/F, i, N)$	$(P/A, i, N)$	$(A/P, i, N)$	$(A/G, i, N)$
1	1.0600	0.9434	1.0000	1.0000	0.9434	1.0600	0.0000
2	1.1236	0.8900	2.0600	0.4854	1.8334	0.5454	0.4854
3	1.1910	0.8396	3.1836	0.3141	2.6730	0.3741	0.9612
4	1.2625	0.7921	4.3746	0.2286	3.4651	0.2886	1.4272
5	1.3382	0.7473	5.6371	0.1774	4.2124	0.2374	1.8836
6	1.4185	0.7050	6.9753	0.1434	4.9173	0.2034	2.3304
7	1.5036	0.6651	8.3938	0.1191	5.5824	0.1791	2.7676
—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—
100	339.3021	0.0029	5638.3681	0.0002	16.6175	0.0602	16.3711

period). If the payment period and the compounding period are equal, then $k = m$, proving that $C = 1$. Then the periodic interest rate is as was calculated above ($i_m = r/m$). However, if the payment period is not the same as the compounding period, then k does not equal m and the following equation must be used to calculate the effective interest rate:

$$i = \left(1 + \frac{r}{M}\right)^c - 1$$

Example 1.

Find the effective interest rate per quarter at a nominal rate of 36% compounded weekly.

Solution:

Identify the items from the word problem and list them with appropriate notation:

$$r = 36\% \text{ (nominal rate)}$$

$$m = 52 \text{ weeks (compounding periods per year)}$$

$$c = 13 \text{ weeks per quarter (number of interest periods per payment periods)}$$

$$\text{Effective rate } i = (1 + 0.36/52)^{13} - 1 = 9.38\% \text{ per quarter.}$$

6 SINGLE SUM AND UNIFORM, GRADIENT, AND GEOMETRIC SERIES

Manipulation of cash flows occurs in a variety of ways. And equivalence provides the concept by which one may adjust cash flows into an equivalent sum or series. The most common manipulation involves converting all units into a single sum in the present at time $N = 0$. This is referred to as the present worth (P). Another common manipulation converts all the cash flows to a single sum at any time $= N$ in the planning horizon. This is called future worth (F). If the cash flows are converted to a single sum at time $N = N$, or the end of the planning horizon, then it is called salvage value (S). However, S is still a function of the future worth. See Fig. 1a.

Another common manipulation is to reduce all cash flows to a uniform series. This cash flow would run from $N = 1$ to $N = N$ at an unchanging dollar amount for every time period. This is referred to as annual worth (A). Here we compare alternatives based on their equivalent annual cash flows. Examples of annual worth (A) would be a car loan, mortgage, or other type of uniform payment over time.

The series that can exist do not always need to be uniform. Gradient series are a conversion series that increases or decreases by a uniform amount G . In the gradient series, the amount increases by G starting with the end of the second period until $N = N$, at which time the cash flow is calculated as $(n - 1)G$. (Note the gradient does not start at time $N = 1$) Gradients are used in replacement studies where the operating and maintenance expenses are assumed to increase by a fixed amount each year. See Fig. 2a.

The last common manipulation is a series that increases or decreases at a constant rate or percentage. This is called a geometric gradient (g) where g increases starting at the end of the second period. Examples of geometric gradients are inflation, growth in sales volume, etc. See Fig. 2b.

All of these conversions, whether together or alone, allow for the representation of most circumstances where money changes hands. This representation could be on the personal or business level. When put on a cash flow diagram, it can be converted, and therefore reduced to one solution. As mentioned with equivalence, the combinations of series and sums can be converted to a single series or sum. This could be done through the use of the conversion factors as presented in Table 4.

Although equivalence allows one to convert from any P to any F , etc., when considering combining and solving cash flows, the following rules need to be kept in mind:

- One can add or subtract negative or positive cash flows if they occur at the same point in time.
- The thing that must be remembered for this to be true, however, is that the cash flows must be the same type of cash flow (i.e., an A cannot be subtracted from a P).

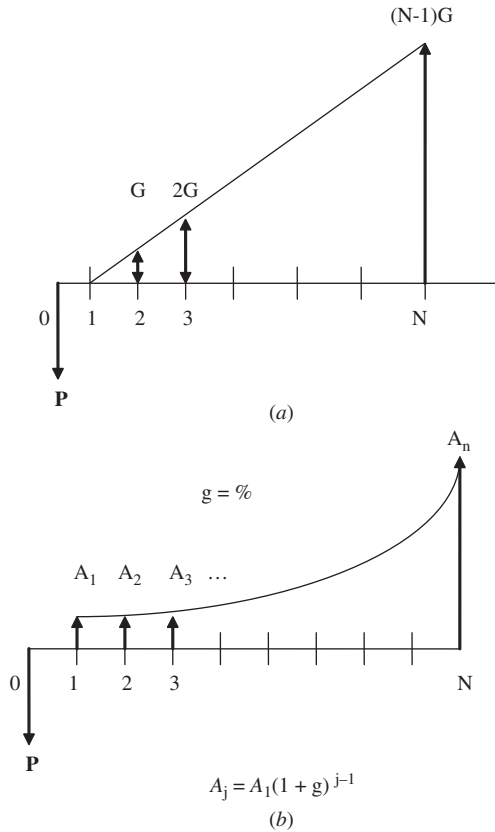


Figure 2 (a) Example of an arithmetic gradient cash flow diagram and (b) example of a geometric gradient cash flow diagram.

Table 4 Conversion of Worth Factors

If you are given →		P	F	A	G
	P	P	P/F	P/A	P/G
To: Solve for an item	F	F/P	F	F/A	F/G
Use factors in Table	A	A/P	A/F	A	A/G

$A = AW =$ Annual worth

$F = FW =$ Future worth

$P = PW =$ Present worth (NPV = net present value)

Note: There is commonly no (P/G) factor. Therefore, to get (P/G) simply multiply $(A/G)(P/A) = (P/G)$.

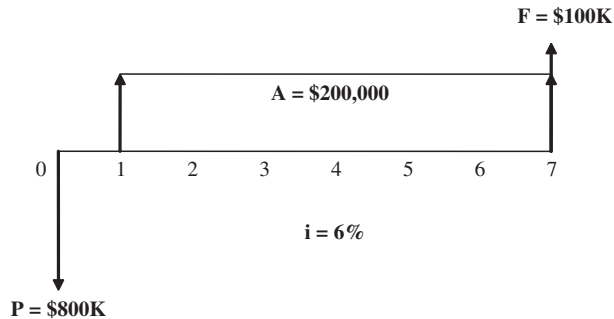
- Planning horizons must be displayed so that they always begin at time $N = 0$. It often is the case that expenses have occurred prior to the start of the new project under consideration. However, these are sunk costs and not considered part of the engineering economic problem unless they have tax consequences.
- Finally uniform or gradient series may need to be manipulated.

Example 2 summarizes the basics of Engineering Economy discussed thus far.

Example 2. Basic Engineering Economics Problem.

A company wants to purchase equipment today for \$800,000. During each year of ownership of the equipment, the company estimates that an additional \$200,000 in revenues will be brought in. The company plans to hold the equipment for 7 years and estimates it will be able to sell the equipment in year 7 for \$100,000. The interest rate is estimated at 6%.

- Draw a cash flow diagram. Identify the items from the word problem and list them with their appropriate notations.
- Solve the problem using the formulas, tables similar to Table 3, and the Excel functions. Is this a profitable enterprise and should it be undertaken?

Cash Flow Diagram

Identify the items from the word problem and list them with their appropriate notations.

$$P = -\$800,000$$

$$A = \$200,000$$

$$S \text{ (also called } F) = \$100,000$$

$$i = 0.06$$

$$N = 7$$

Solution

Solve the problem using the equations in Table 2.

$$\begin{aligned}
 PW &= -P + A \left(\frac{P}{A}, i, N \right) + F \left(\frac{P}{F}, i, N \right) \\
 &= -800,000 + 200,000 \left(\frac{P}{A}, 6\%, 7 \right) + 100,000 \left(\frac{P}{F}, 6\%, 7 \right) \\
 &= -800,000 + 200,000 \left[\frac{(1+i)^N - 1}{i(1+i)^N} \right] + 100,000 \left[\frac{1}{(1+i)^N} \right] \\
 &= -800,000 + 200,000 \left[\frac{(1+.06)^7 - 1}{.06(1+.06)^7} \right] + 100,000 \left[\frac{1}{(1+.06)^7} \right] \\
 &= -800,000 + 200,000 (5.5824) + 100,000 (.6651) \\
 &= -800,000 + 1,116,480 + 66,510 = \$382,990
 \end{aligned}$$

Yes, this would be a profitable project that should be undertaken since the present worth of the project is positive.

Solution

Solve the problem using the factors from the interest tables such as Table 3.

$$\begin{aligned} PW &= -P + A (P/A, i, N) + F (P/F, i, N) \\ &= -800,000 + 200,000 (P/A, 6\%, 7) + 100,000 (P/F, 6\%, 7) \\ &= -800,000 + 200,000 (5.5824) + 100,000 (.6651) \\ &= \$382,990 \end{aligned}$$

Yes, this would be a profitable project that should be undertaken since the present worth of the project is positive.

$$\begin{aligned} PV &= (\text{rate, nper, pmt, fv, type}) \\ &= (.06, 7, -200000, -100000, 0) - 800,000 = \$382,982 \end{aligned}$$

Yes, this would be a profitable project that should be undertaken since the present worth of the project is positive.

7 COMPARING ALTERNATIVES: DEFINING ALTERNATIVES

This chapter on engineering economics started out by asking which method one would choose to get commuters from one side of the Hudson River to the other. That is just one example of an engineering economics question whose root is to find the most economical solution among several alternatives. One could argue in this globally connected world, that the key to global competitive advantage is the successful implementation of engineering design. In particular, the ability to design for cost or design for affordability being the difference between producing profits and going out of business. Today there are many, many ideas out there for new apps, new businesses, or just new technologies. But, there are limited dollars that can be invested into making these ideas a reality. How does one decide in which project or design to invest? How does one rank order to a set of alternatives? The answer is simply engineering economics. Through financial criteria called ‘figures of merit’, engineering economics provides a methodical way to rank and compare investment opportunities and make business decisions based on their financial impact.

As described at the start of this chapter, there are often both technical and financial considerations to design concepts. To pick the ideal solution, or set of solutions, both concepts need to be considered. For technical feasibility, one might consider the example of the alternative fuel vehicle discussed earlier, where weight of material or size of the fuel tank are questions integral to the fuel efficiency of the vehicle. For financial feasibility, one must consider the amount of funds available. In some opportunities it might be necessary to consider both feasibilities at the same time. For example, maybe the scope of the design makes for it to have an alpha version in year 1 with some basic functions, while the beta version produced in year 2 will be the full fledged model with all the bells and whistles incorporated into the design. In some cases even the “do nothing” alternative may need to be considered in the set of alternatives since it might be better to pass on some alternatives entirely.

Lastly, in addition to the concepts that must be considered above, while choosing among alternatives an interest rate or level of risk must be chosen. This rate is known as the MARR (minimal attractive rate of return). MARR is simply the rate of return that an organization feels it must get in order for a project to be attractive to pursue. This rate is normally chosen based on the level of risk associated with the undertaking as well as the current interest rate, inflation,

etc. Thus, the higher these factors, the higher the MARR, or minimum interest rate that will be desired.

8 COMPARING ALTERNATIVES THROUGH THE FIGURES OF MERIT

There are generally five figures of merit used to compare and rank alternative investment opportunities. They are:

- Present value, present worth, or net present worth (PV, PW, or NPW)
- Future value, future worth, or net future worth (FV, FW, or NFW)
- Annual worth or annual equivalence (AW or AE), or if the annual worth is negative then annual cost (AC) or equivalent uniform annual cost (EUAC)
- Internal rate of return (IRR)
- Benefit–cost analysis (BCA) Or Benefit Cost Ratio (BCR)

The first three items are called the “three worths” and are directly linked though the rate of return factors. As previously mentioned, equivalence allows us to convert from any worth to any other worth through these return factors (see Table 4). The last two items, IRR and BCA cannot be compared directly or converted into each other, as they do not take the magnitude of the investment into account.

8.1 Present Value

Present value is the most widely used figure of merit. It is equal to the discounted value of all future sums and series when discounted at a given rate. The PV is commonly used since it resolves all future investments in terms of the present time; something that is easy for most people to compare. The PW might be a valuable figure of merit when comparing investment in equipment purchases from various manufacturers. In cases when one is trying to maximize the benefits, one would want the PV to be as high as possible. However, if one were looking at a project where only costs were involved, then one would want to minimize the costs, or choose the project with the lowest present value. Example 2 provides a solution to solving for the present value using mathematical calculations and Excel.

8.2 Future Value

The future value (FV) of an asset is the value of that asset at a specified date in the future that is equivalent to a specified sum today. Future worth is a compounded value that can be obtained by multiplying the present or annual worth by the proper F/P or F/A factors. The future value of an investment may be the best figure of merit in the case of reviewing potential IRA distributions upon retirement. As was the case with present value, when trying to maximize the benefits, the alternative with the highest future value is chosen. If one is dealing only with costs, then the alternative with the lowest value, or least cost, is chosen. Please see Example 3.

8.3 Annual Worth

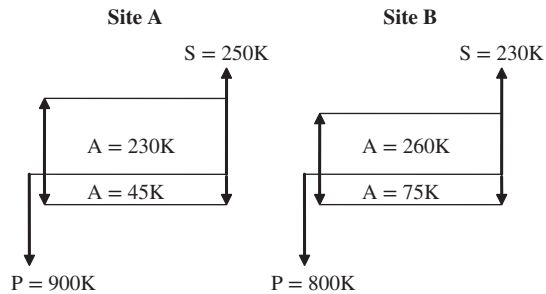
The annual worth (AW) of an asset is the equivalent uniform cost/revenue incurred/gained over the entire planning horizon of the project. It is the sum of all present and future cash transactions converted to a uniform series. Annual worth is a valuable figure of merit when looking at car loans or home mortgages. As is the case with all worths, when one is trying to maximize the benefits, the highest annual worth alternative is chosen. If the problem is one of only costs, then the alternative with the lowest value, or least cost, is chosen. Please see Example 3.

Example 3.

Connolly Construction Company is planning to invest in a new yard in which to keep their equipment. The economics of the two alternatives are provided below.

Description		Location A	Location B
First cost	P	\$900,000	\$800,000
Annual revenue	$+A$	\$230,000	\$260,000
Annual expense	$-A$	\$45,000	\$75,000
Salvage value	S	\$250,000	\$230,000
Life	N	12	12
MARR	i	10%	10%

- a. Draw the cash flow diagrams for the two locations:



- b. Calculate the present worth (PW) for both locations:

$$PW = A(P/A, i, N) + S(P/F, i, N) - P$$

$$PW \text{ site A} = (230,000 - 45,000) \times (P/A, 10\%, 12) + 250,000(P/F, 10\%, 12) - 900,000$$

$$PW \text{ site A} = (230,000 - 45,000) \times (6.8137) + 250,000 \times (0.3186) - 900,000$$

$$PW \text{ site A} = \$440,191$$

$$PW \text{ site B} = (260,000 - 75,000) \times (P/A, 10\%, 12) + 230,000(P/F, 10\%, 12) - 800,000$$

$$PW \text{ site B} = (260,000 - 75,000) \times (6.8137) + 230,000 \times (0.3186) - 800,000$$

$$PW \text{ site B} = \$533,818$$

- c. Calculate the future worth (FW) for both locations.

$$FW = S + A(F/A, i, N) - P(F/P, i, N)$$

$$FW \text{ site A} = 250,000 + 185,000(F/A, 10\%, 12) - 900,000(F/P, 10\%, 12)$$

$$FW \text{ site A} = 250,000 + 185,000(21.3843) - 900,000(3.1384)$$

$$FW \text{ site A} = 1,381,505$$

Or more simply

$$FW \text{ site A} = P(F/P, i, N) = 440,191(3.1384) = \$1,381,495$$

Or in Excel

$$FW = FV(\text{rate}, \text{nper}, \text{pmt}, \text{pv}, \text{type}) + 250,000$$

$$FVW = FV(.10, 12, -185000, 900000, 0) + 250,000 = 131,506$$

Note small error from \$1,381,505 due to rounding:

$$\begin{aligned} FW \text{ site B} &= 230,000 + 185,000(F/A, 10\%, 12) - 800,000(F/P, 10\%, 12) \\ &= 230,000 + 185,000(21.3843) - 800,000(3.1384) \\ &= \$1,675,348 \end{aligned}$$

Or more simply

$$FW \text{ site B} = P(F/P, i, N) = 533,818 (3.1384) = \$1,675,334$$

Or in Excel

$$FW = FV(\text{rate}, \text{nper}, \text{pmt}, \text{pv}, \text{type}) + 250,000$$

$$FVW = FV(.10, 12, -185000, 800000, 0) + 230,000 = 1,675,349$$

Note small error from \$1,675,348 due to rounding.

- d. Calculate the annual worth (AW) for both sites.

$$AW = A + S(A/F, i, N) - P(A/P, i, N)$$

$$\begin{aligned} AW \text{ site A} &= (230,000 - 45,000) + 250,000(A/F, 10\%, 12) - 900,000 \\ &\quad (A/P, 10\%, 12) \\ &= (230,000 - 45,000) + 250,000(0.0468) - 900,000(0.1468) \\ &= \$64,604 \end{aligned}$$

Or more simply

$$AW \text{ site A} = P(A/P, i, N) = 440,191 (0.1468) = \$64,620$$

Or in Excel

$$PMT = FV(\text{rate}, \text{nper}, \text{pv}, \text{fv}, \text{type}) + 185,000$$

$$PMT = FV(.10, 12, 900000, -250000, 0) + 185,000 = \$64,603$$

Note small error from \$64,604 due to rounding:

$$\begin{aligned} AW \text{ site B} &= (260,000 - 75,000) + 230,000(A/F, 10\%, 12) - 800,000 \\ &\quad (A/P, 10\%, 12) \\ &= (260,000 - 75,000) + 230,000(0.0468) - 800,000(0.1468) \\ &= \$78,345 \end{aligned}$$

Or more simply

$$AW \text{ site B} = P(A/P, i, N) = 533,818 (0.1468) = \$78,364$$

Or in Excel

$$PMT = FV(\text{rate}, \text{nper}, \text{pv}, \text{fv}, \text{type}) + 185,000$$

$$= FV(.10, 12, 800000, -230000, 0) + 185,000 = \$78,345$$

Note small error from \$78,345 due to rounding.

- e. Which alternative should be chosen choose? Why?
Choose site B because it has a much greater/higher group of values or “worths.”

8.4 Internal Rate of Return

The internal rate of return is defined as the rate of return that makes the net present worth of a set of cash flow equal to zero. The interest rate that makes this zero is termed as the internal rate of return.

$$\Sigma PW = 0$$

The internal rate of return is the return investors could expect if the cash flow estimates in the cost studies prove true. It is also referred to as the cutoff rate, targeted rate, hurdle rate, or profitability index. These descriptive names are suitable as the IRR is the estimated “target” earning rate for the project.

Often this target rate is compared to the MARR in order to determine if a project should be undertaken. The rules for this comparison are as follows: If the MARR is the minimum acceptable rate at which a return on an investment would be considered favorable, then any investment for which the return is greater than MARR would be considered desirable. And any investment that was below MARR would not be considered desirable.

Simply,

- If $IRR > MARR$, the investment is worthwhile.
- If $IRR < MARR$, the investment should not be undertaken.

There are two negative aspects to IRR. The first is that multiple solutions can be derived. This occurs when there is a large capital investment in later years, where the cash flow can move from positive to negative in those later years. This may result in multiple changes in sign and therefore multiple solutions. The second aspect is that IRRs from different projects cannot be compared directly since IRR does not take the magnitude of the investment into consideration. For example, consider an IRR of 25% on an investment of a few *hundred* dollars versus an IRR of 20% on an investment of a few *billion* dollars. The two are apples and oranges and cannot be compared directly. When one wants to compare projects using IRR as the figure of merit, incremental analysis must be completed on each set of projects and the following rule must be used to determine which project is best in each comparison.

If $IRR_{\text{Beta-Alpha}} > MARR$, then beta is best choice.

If $IRR_{\text{Beta-Alpha}} < MARR$, then alpha is best choice.

Please note, the way to determine if one should subtract beta minus alpha, or alpha minus beta, is to always have the incremental P as a negative value. See Example 4 for details on how to solve an incremental IRR problem.

Example 4.

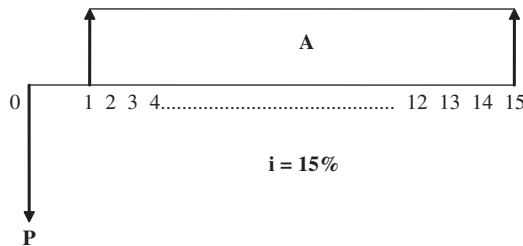
Mr. Rogers is planning to invest in a new property that he will rent out for income. He is considering two alternatives, and the economic details of both of them are listed below.

- a. Draw the generic cash flow diagram for the opportunities involved.
- b. What is the best project using MARR?
- c. What is the best project using IRR?

Description		Apartment Building A	Apartment Building B
First cost	P	\$475,000	\$575,000
Annual income	$+A$	\$100,000	\$125,000
Annual expense	$-A$	\$20,000	\$40,000
Salvage value	S	\$330,000	\$380,000
Life	N	15	15
MARR	i	8%	8%
IRR	i^*	17%	14%

Solution

a. The generic cash flow diagram appears below:



b. What is the best project using MARR?

To solve this problem, calculate the NPV, also called the present worth for each alternative:

$$NPV = -P + [A_{\text{revenues}} - A_{\text{expenses}}](P/A, i, N) + S(P/F, i, N)$$

$$NPV_A = -475,000 + [100,000 - 20,000](P/A, 8\%, 15) + 330,000(P/F, 8\%, 15)$$

$$NPV_A = -475,000 + 80,000(8.5595) + 330,000(.3152) = \$313,788$$

$$NPV_B = -575,000 + [125,000 - 40,000](P/A, 8\%, 15) + 380,000(P/F, 8\%, 15)$$

$$NPV_B = -575,000 + 85,000(8.5595) + 380,000(.3152) = \$272,348$$

Thus, pick apartment building A since it has the greatest NPV.

c. What is the best project using IRR?

IRRs cannot be compared directly. Thus, the information provided in the problem statement about IRR is immaterial. One must use the same NPV equation as used in part (a) to solve for IRR, only this time we set NPV equal to zero as necessary and explained above:

$$NPV = -P + [A_{\text{revenues}} - A_{\text{expenses}}](P/A, i, N) + S(P/F, i, N)$$

*Note: all values below in thousands

$$NPV_{B-A} = 0 = -575 - (-475) + (85 - 80)(P/A, i^*, 15) + (380 - 330)(P/F, i^*, 15)$$

$$\text{First guess} = -100 + 5(P/A, 1\%, 15) + 50(P/F, 1\%, 15) = \$12.390$$

$$\text{Second guess} = -100 + 5(P/A, 3\%, 15) + 50(P/F, 3\%, 15) = -\$8.215$$

Solution Table

Rate (%)	NPV
1%	\$12.390
i	\$0
3%	-\$8.215

Interpolate the solution for IRR using the numbers from the Solution Table:

$$\frac{1\% - i^*}{1\% - 3\%} = \frac{\$12.390}{\$12.390 - \$8.215}$$

The answer is $i^* = \text{IRR} = 2.2\%$

The rule states that if the $\text{IRR}_{\text{Beta-Alpha}} < \text{MARR}$, then alpha is the best choice. Thus, choose apartment building A since $\text{IRR} < \text{MARR}$.

8.5 Benefit–Cost Analysis

The benefit–cost ratio (BCR) is the last figure of merit. For projects aimed at improving the welfare of the public, BCR is often the preferred figure of merit. The reason for this is that the private sector is focused on revenues, while the public sector measures things in terms of benefits to the community, and whether these benefits outweigh the costs. This is the figure often used by government entities to justify their selection of projects from the normally very large pool of projects seeking funding.

The notion of the BCR finds its history in the Flood Control Act of 1936. This act stated that “the benefits to whomsoever they accrue are in excess of the estimated costs.” This was a way for the government to specify that a project would be considered worthwhile if the benefits derived from the project exceeded the estimated costs of the project. What was happening in 1936 played a role in setting this overall general objective. The economy was in a terrible state and the United State was in the midst of the Great Depression. The response of the U.S. government was to fund many capital projects to “pump” money into the economy. How the government decided which projects to fund, and which ones not to fund, was through the concept that the benefits must exceed the costs. To convert this to mathematical terms, the following would be the rules that must be observed:

$$B > C$$

$$B - C > 0 \text{ or } B/C > 1$$

All preceding equations mean the same thing and lead to capital efficiency, allocating money to worthwhile projects that will provide the most benefit.

$$B/C \text{ Ratio} = \frac{\text{PW benefits}}{\text{PW costs}} = \frac{\text{FW benefits}}{\text{FW costs}} = \frac{\text{AW benefits}}{\text{FW costs}}$$

The above equation can be further refined into conventional and modified formulas. The difference between the two formulas lies in the placement of the operating and maintenance (O&M) costs: either in the denominator (for the case of the conventional) or in the numerator (for the case of the modified). Although the two equations will produce two numerically different answers, both equations will provide the same recommendation based on the rules stated above:

$$\begin{aligned} \text{Conventional } B/C \text{ Ratio} &= \frac{\text{Worth of the benefits}}{\text{Worth of the initial costs} + \text{Worth of the O\&M costs}} \\ \text{Modified } B/C \text{ Ratio} &= \frac{\text{Worth of benefits} - \text{Worth of the O\&M costs}}{\text{Worth of the initial costs}} \end{aligned}$$

It must be remembered that all benefits and costs must either be brought to the present or annualized. Oftentimes the initial cost is in terms of P , but the ongoing costs are in terms of A . Prior to calculating the BCR, all items must be annualized or brought to the present value. See Example 5 for details. Additionally, just as the IRR of individual projects can be calculated, so too can the individual BCRs of projects be calculated. The first step when comparing alternatives is to calculate the BCR for each project. And just like the IRR does not capture the magnitude of the investment, so too does BCR not capture the magnitude of the investment. Therefore like IRR, when comparing alternatives, incremental analysis must also be used in determining the best project using BCR. The rule for comparing alternatives using incremental analysis is as follows:

$$\text{If } B/C_{X-Y} > 1, \text{ choose } Y$$

$$\text{If } B/C_{X-Y} < 1, \text{ choose } X$$

Example 5.

A municipality wants to build a new highway to ease congestion on the current highway and reduce travel time.

Initial cost of expansion = \$1,000,000

Annual savings through benefits to travelers = \$135,000

Annual operating and maintenance costs = \$50,000

Value at end of useful life = \$400,000

Useful life = 30 years

Interest rate = 5%

$$\text{Conventional BCR} = \frac{135,000 (P/A, 5\%, 30) + 400,000 (P/A, 5\%, 30)}{1,000,000 + 50,000 (P/A, 5\%, 30)}$$

$$\text{Modified BCR} = \frac{135,000 (P/A, 5\%, 30) + 400,000(P/F, 5\%, 30)}{1,000,000 - 50,000(P/A, 5\%, 30)} = 1.4$$

9 CAPITAL RECOVERY, CAPITALIZED COST, AND REPLACEMENT STUDIES

9.1 Capital Recovery

Capital recovery (CR) of an investment is a uniform series representing the difference between the annualized equivalent of the first cost and the annualized equivalent of the salvage value. This is necessary since the capital costs are normally nonrecurring, but the operating costs over the life of the asset are normally estimated on an annual basis. Therefore, one must translate this one-time (purchase minus salvage value) cost over the life of the project. The capital recovery equation is valuable since an organization can compare the additional estimated annual revenues from the purchase of equipment to its total annual equivalent cost (total AE).

- If the revenues are larger, than the machine should be purchased.
- If the total AE is larger, than the machine will cost more money than it can produce, thus the recommendation would be to not purchase the new equipment.

$$\text{CR} = P(A/P, i, N) - S(A/F, i, N)$$

$$\text{Total AE} = \text{CR} + \text{O\&M}$$

9.2 Capitalized Cost

Capitalized cost (CC) is a special case of present worth. It is a concept used when the useful life of an asset is estimated to be very long, normally greater than 40 years. When this occurs, the assumption can be made that N is infinite. This kind of infinite assumption can be easily visualized through governmental buildings, roadway pipelines, etc. where these items are basically considered permanent. The capitalized cost is the funds that would need to be set aside now, at some interest rate, to yield the funds required to take care of the asset indefinitely. In this case, the interest is spent, but the principal remains untouched.

$$CC = P = A/i$$

9.3 Replacement Studies

Up to this point we have considered the purchase or selection of new items. However, oftentimes one wants to replace an existing asset with the same or similar asset. For most engineers the decision is not to build a new facility but rather how to allow the current facility to operate in the most economical way possible. So the question becomes when should equipment be replaced. In Engineering Economics this is called a replacement study, and the concept of comparing the defender (existing machine) and challenger (the proposed replacement) begins.

The defender normally has only a few years left (if that) of useful life. The challenger, being new, will have many years of useful life ahead. This means the two have unequal lives, making them difficult to compare directly. As such the annual marginal costs and EUAC values are items examined to determine if the asset should be replaced or kept for another year.

Replacement analysis can take up a whole chapter in a textbook and is therefore beyond the scope of this handbook. For more information the reader is referred to Lang and Merino.⁴

10 ADDITIONAL ANALYSES AND CONSIDERATIONS IN THE SELECTION PROCESS

10.1 Depreciation

Depreciation is a method organizations use to distribute the cost of an asset over a long period of time. As an example, consider a company investing in a large piece of equipment. By U.S. tax law, that company must allocate the cost of that equipment over the span of a few years. If the company did not do this, then the financial statements of the company in the year they bought the equipment would be significantly worse than for other years. So depreciation can be defined as a noncash expense that allocates the cost of a capital asset over its useful life.

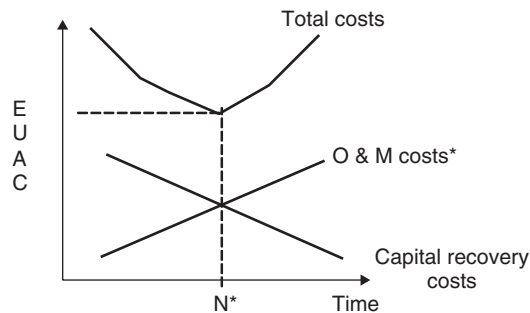


Figure 3

The U.S. government also defines the useful life of many assets. The tax authorities allow several depreciation methods to be used. In this text we will describe straight line (SL), double declining balance (DDB), and MACRS.

Straight-line depreciation assigns an equal amount of depreciation expense to each year of the asset's useful life. The depreciable cost of the asset is divided by the useful life to determine the annual depreciation expense. The formulas for the straight-line depreciation and the straight-line depreciation rate are below:

$$\text{Straight-line depreciation per year} = \frac{\text{First cost-salvage value}}{\text{Useful life in years}}$$

$$\text{Straight-line depreciation rate} = \frac{1}{\text{Useful life in years}}$$

The double declining balance method is an accelerated depreciation method. What this means is that larger amounts are depreciated in the earlier periods of the useful life and smaller amounts are depreciated in the later periods of the useful life. It calculates the annual depreciation by multiplying the asset's book value by a constant rate, which is two times the straight-line depreciation rate. Hence, the name *double* declining balance method.

$$\text{Double Declining Balance Rate per Year} = 2 * \text{Straight-Line Depreciation Rate}$$

MACRS stands for modified accelerated cost recovery system. It is also an accelerated depreciation system that allows larger amounts to be depreciated earlier in the useful life and smaller amounts later on. MACRS is mandatory for federal tax returns under the 1986 Tax Reform Act. MACRS defines eight property classes, and the depreciation of a particular asset depends in which classification the asset falls. To determine the annual depreciation expense of an asset using MACRS, one must first locate the asset in the appropriate property class table. In the table find the years of ownership allowed, and multiply the first cost of the asset by the percentage given in the table for year 1. For year 2, simply multiply the first cost by the percentage given in the table for year 2, etc, etc.

The differences between the three kinds of depreciation are graphically represented in Fig. 4.

10.2 Inflation

Inflation and purchasing power are inversely related. As inflation goes up, purchasing power, or the amount one can purchase for a specified sum of money, goes down. These concepts do not just increase, or just decrease, but rather fluctuate in all the world's economies. In terms of engineering economics, inflation makes future dollars less valuable than present dollars. Since engineering economics requires equivalence, it is important for us to be able to incorporate

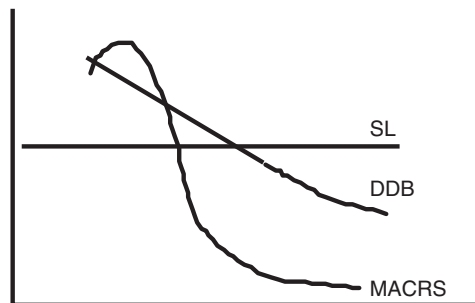


Figure 4 Depreciation rates.

concepts such as inflation into our analyses. The three elements to review when considering inflation are listed below, along with the mathematical relationship between them:

- Market interest rate (i). This is the rate of interest in the general marketplace. It is also referred to as MARR, or MARR with inflation included within it.
- Real interest rate (i'). This rate is the real rate of the growth of money excluding the effect of inflation. It is also referred to as MARR without inflation.
- Inflation rate (f). This is the rate of increase in the number of dollars needed to purchase the same good:

$$i = i' + f + i' * f$$

There are also two different types of dollars to be considered when discussing inflation. They are listed below along with the relationship between interest rate and types of dollars.

- Actual Dollars. This is the actual amount of the dollar, meaning that inflation is already a component of the dollar. As such when using these dollars, one needs to use i , the market interest rate.
- Real Dollars. These are the “constant” dollars as they do not have the effect of inflation in them. They are sometimes called “constant purchasing power dollars.” As such when using these dollars, one needs to use the i' , the real interest rate.

10.3 After Tax Analysis

Discussions on resolving cash flows up to this point have been on a before-tax basis. But often taxes are a large sum that can impact cash flow. Although the consideration of after tax makes for a more elaborate calculation, the use of excel can stream line the process. Of all the financial analyses conducted in an organization, the after tax analysis can be the most important. Taxes paid and tax-related items, like depreciation, have a significant impact on capital investment opportunities. There are seven key steps to completing an after tax analysis:

1. Identify the projects under consideration that fall within the strategy of the organization.
2. Evaluate the relevant mutually exclusive alternatives.
3. Estimate the income statement for each alternative.
4. Estimate the balance sheet for each alternative.
5. Conduct the economic analysis for each alternative using the information from 3 and 4
6. Choose the most economical alternative. Review if noneconomic factors should influence choice as well.
7. Include the selected projects in the capital budget of the organization.

After tax analysis can take up several chapters in a textbook and are thus beyond the scope of this handbook. For more information the reader is referred to Ref. 4.

10.4 Sensitivity Analysis

Engineering Economic studies conducted are often regarding potential investments to be made today or over some time period. However, they also take into account costs or revenues that may occur in the future from investments made today. As these future numbers are estimates, and can therefore vary given changes in the economy, etc., knowing which estimates are the most sensitive to change is valuable knowledge. Sensitivity analysis provides one with the information to know which items involved in the figure of merit calculation vary widely, and

which will remain relatively constant. By knowing which items vary greatly, one knows which items to focus on, and which to not be too concerned about. By knowing which items are the most sensitive, one can examine more closely the details of these items, and the impact of their change in order to have greater confidence in the calculated figure of merit.

The first step in sensitivity analysis is to calculate the base case. This would simply be calculating the present worth (or any other figure of merit) as instructed to do thus far, using the most likely values for the inputs. Next, change one variable by some percentage above and below that used in the base case while holding the remaining variables constant. Then calculate the new present worth using the new values. Results can be displayed both graphically and in tabular form. Graphically, the variable(s) that had the greatest slope(s) would be the most sensitive variable(s). In a tabular calculation, the item(s) that had the largest numerical value(s) as its sensitivity would be the most sensitive item(s). The most sensitive item is the variable that would affect the figure of merit the most given any change to the variable.

10.5 Break-Even Analysis

Break-even analysis is a useful tool when determining if one should proceed with an investment into a product, for example. Break-even analysis calculates the point at which if you go beyond it, profit is to be made, or if you fall short of it, there will be a loss. For example, let us say a company wanted to produce autonomous underwater vehicles. They determined that using the specifications provided, 3000 units would need to be sold each year to “break even” with the investment of producing the units. The company would instantly know from this calculation that it should scrap the design, since there are not even 3000 underwater laboratories in the world, let alone the desire for each of them to buy a new underwater vehicle each year. If, however, the company was looking at producing laser pointers and the break-even point calculated for the specifications of the laser pointer design was 1500, then the company would know to move forward. The company could easily sell 1500 pointers and many more, ensuring a profit for the company.

Break-even analysis can also be used in a lease-versus-buy solution. In this type of analysis, the costs associated with purchasing the vehicle would be compared with those of leasing the vehicle. The point calculated would be that point at which it is more or less expensive to lease than to buy. This type of analysis would be used when only a single factor is at play, such as utilization of the equipment.

10.6 Risk Analysis

Risk analysis is used when there is a lack of definite or precise knowledge regarding a future occurrence or condition. Things are at risk because there is a chance of getting something other than the expected value. In these cases, the probabilities of the risk are estimated, and decision rules can be used to assist in decision making. Some of these rules include: most probable future value, dominance and aspiration level, and laws of expected value and variance. In addition, decision trees also address risk in the calculation of their outcomes. Lastly, more complicated calculations such as modeling and simulations can be used to imitate real-life situations and their associated risk profiles.

The details of risk analysis, as well as details on sensitivity and break-even analyses, are beyond the scope of this handbook. For more information the reader is referred to Newnan et al.¹

11 CONCLUSION

Engineering Economy is the study of the feasibility, and evaluation of the cost, of possible solutions to engineering problems. Since Engineering Economics is applicable to all disciplines of engineering, it is a required section of the fundamentals in engineering (FE) exam as well as the professional engineering (PE) exam, necessary to attain a professional license.

Engineers faced with design decisions that deal with multiple designs, methods or materials commonly encounter the topic of engineering economics. The choices made in choosing these ultimately affect the majority of costs for manufacturing or construction. Therefore, besides meeting the technical scope of a design, the alternative solutions must also be economically viable. An engineer often chooses the requirements for the design. Therefore, the design alternatives chosen and the comparison of them can maximize benefits or minimize cost, indicate which projects can be deemed worthwhile and which are not, and which should be given higher priority simply based on potential economic impact.

It is for these reasons that Arthur Wellington⁵ defined engineering as “the art of doing well with one dollar what any bungler can do with two.”

REFERENCES

1. D. G. Newnan, T. G. Eschenbach, and J. P. Lavelle, *Engineering Economic Analysis*, 11th ed., Oxford University Press, New York, 2012.
2. Oxford University Press, *Compound Interest Tables*, accessed July 27, 2012, available: www.oup.com/pdf/ca/compoundstudent.pdf.
3. C. S. Park, *Contemporary Engineering Economics*, 5th ed, Pearson Prentice Hall, Upper Saddle River, NJ, 2011.
4. H. Lang, and D. Merino, *The Selection Process for Capital Projects*, 6th ed., Wiley, Hoboken, NJ, 2002.
5. A.M. Wellington, *The Economic Theory of the Location of Railways: An Analysis*, Wiley, New York, 1887.

CHAPTER 20

EVALUATING AND SELECTING TECHNOLOGY-BASED PROJECTS

Hans J. Thamhain
Bentley University
Waltham, Massachusetts

1 MANAGEMENT PERSPECTIVE	605	2.6 Going Beyond Simple Formulas	611
2 QUANTITATIVE APPROACHES	607	3 QUALITATIVE APPROACHES	611
2.1 Net Present Value Comparison	608	3.1 Collective Multifunctional Evaluations	612
2.2 Return on Investment Comparison	609	4 RECOMMENDATIONS	613
2.3 Cost–Benefit	610	5 TERMS	615
2.4 Payback Period Comparison	610	6 VARIABLES AND ABBREVIATIONS	616
2.5 Pacifico and Sobelman Project Ratings	610	REFERENCES	616

Predicting project success is difficult and often unreliable.^{1–4} The long list of prominent project failures, ranging from computers to pharmaceutical and supersonic transport, reminds us of this reality.^{5,6} Many projects do not live up to their expectations or outright fail even before their technical completion.⁷ The ability to evaluate project proposals, assessing future success and organizational value, is critical to overall business performance for most enterprises.

1 MANAGEMENT PERSPECTIVE

Project success is multifaceted. Typically, it includes not only technical but also financial, marketing, and social, legal, and ethical dimensions. For most projects, the DNA of success is highly complex, and outcomes are difficult to predict.⁸ Evaluating and selecting projects is both an art and a science that, for most cases, has to go beyond a simple cost–benefit analysis. To be sure, few decisions have more impact on business performance than the resource allocations for new projects. Virtually every organization selects and implements projects, ranging from product development to organizational improvements, and from customer contracts to R&D activities and bid proposals. Pursuing the “wrong” project not only wastes company resources but also causes the enterprise to (i) miss critical alternatives, (ii) operate less flexible and responsive in the marketplace, and (iii) miss opportunities for leveraging core competencies. Project opportunities must be analyzed relative to their potential value, strength, and importance to the enterprise. Four major dimensions should be considered: (1) added value of the new project, (2) cost of the project, (3) readiness of the enterprise to execute the project,

and (4) managerial desire. A well-organized project evaluation and selection process provides the framework for systematic data gathering and informed decision making toward resource allocation.^{9,10} Typically, these decisions can be broken into four principal categories:

1. *Deciding Initial Feasibility*: Screening and filtering, quick decision on the viability of an emerging project for further evaluation
2. *Deciding Strategic Value to Enterprise*: Identifying alternatives and options to proposed project
3. *Deciding Detailed Feasibility*: Determining the chances of success for a proposed project
4. *Deciding Project Go/No-Go*: Committing resources for a project implementation

While making these decisions looks simple, logical, and straightforward, developing meaningful support data is a complex undertaking. It is also expensive, time consuming, and often highly eclectic. Typically, decision-making requires the following inputs:

1. Specific resource requirements
2. Specific implementation risks
3. Specific benefits (economics, technology, markets, etc.)
4. Benchmarking and comparative analysis
5. Strategic perspective, including long- and short-term value assessment

Although it is challenging to estimate costs, schedules, risks, and benefits, such as those shown in Table 1, these measures are relatively straightforward in comparison to predicting

Table 1 Typical Criteria and Measures used for Project Evaluation

The criteria relevant to the evaluation and selection of a particular project depend on the specific project type and business situation such as for a particular product development, custom project, process development, industry, or market. Typically, evaluation procedures include the following criteria and measures:

- Development cost
 - Development time
 - Technical complexity and feasibility
 - Risk
 - Return on investment
 - Cost–benefit
 - Product life cycle
 - Sales volume
 - Market share
 - Project business follow-on
 - Organizational readiness and strength
 - Consistency with business plan
 - Resource availability
 - Cash flow, revenue, and profit
 - Impact on other business activities
-

Note: Each criteria is based on a complex set of parameters and variables.

project success. The difficulty is in defining a meaningful aggregate indicator for project value and success. Methods for determining success range from purely intuitive to highly analytical. No method is seen as truly reliable in predicting success, especially for more complex and technologically intensive types of projects. Yet, some companies have a better track record in selecting “winning” projects than others. They seem to have the ability to create a more integrated picture of the potential benefits, costs, and risks for the proposed project relative to the company’s strength and strategic objectives. Producing such a composite is both a science and an art. Traditionally, the management literature suggested by-and-large rational selection processes to support project selections.¹¹ However, purely rational-analytical processes apply only to a limited number of business situations. Many of today’s technologically complex business scenarios require the integration of both analytical and judgmental techniques to evaluate projects in a meaningful way,^{12,13} predicting success and making the best choice. Yet, in spite of the dynamics involved in the selection process, systematic information gathering and standardized methods are at the heart of any project evaluation process and provide the best assurance for reliably predicting project outcome and repeatability of the decision process. Approaches to project evaluation and selection fall into one of three principal classes:

1. Primarily *quantitative* and *rational* approaches
2. Primarily *qualitative* and *intuitive* approaches
3. Mixed approaches, combining both quantitative and qualitative methods

Because of the interdisciplinary complexities involved, analyzing a new project opportunity is a highly interactive effort among the various resource groups of the enterprise and its partners.¹⁴ Often, many meetings are needed before (i) a clear picture emerges of potential benefits, costs, and risks involved in the project, and (ii) data emerge that is useful for the project evaluation and selection process, regardless of its quantitative, qualitative, or combined nature.

2 QUANTITATIVE APPROACHES

Quantitative approaches are often favored to support project evaluation and selections if the decisions require economic justification. They are also commonly used to support judgment-based project selections. One of the features of quantitative approaches is the generation of numeric measures for simple and effective comparison, ranking, and selection.^{11,15} These approaches also help to establish quantifiable norms and standards and lead to repeatable processes. Yet, the ultimate usefulness of these methods depends on the assumption that the decision parameters, such as cash flow, risks, and the underlying economic, social, political, and market factors, can actually be quantified and reliably estimated over the project life cycle. Therefore, quantitative techniques are effective and powerful decision support tools, if meaningful estimates of cost–benefits, such as capital expenditures and future revenues, can be obtained and converted into net present values for comparison. Because of their importance, quantitative methods have been discussed in the literature extensively, ranging from simple return on investment calculations to elaborate simulations of project scenarios. Many companies eventually develop their own project evaluation/selection models, customized to their specific needs. However, the backbone for most of these customized models is a set of economic/financial measures that tries to determine the cost–benefit of the proposed venture, usually for some point in the future. Specifically, four measures are especially popular:

1. Net present value (NPV)
2. Return on investment (ROI)
3. Cost–benefit (CB)
4. Payback period (PBP)

The calculation and application of these measures to project evaluation/selection will be illustrated by case examples. Specifically, four project proposals (described in Table 2) are evaluated in this chapter, using the above measures. The results are summarized in Table 3.

2.1 Net Present Value Comparison

This method uses discounted cash flow as the basis for comparing the relative merit of alternative project opportunities. It assumes that all investment costs and revenues are known and that economic analysis is a valid basis for project selection.

Table 2 Description of four project proposals

Project option P1: Management does not accept any new project proposal. Hence, no investment capital is required, nor is any revenue generated.

Project option P2: This opportunity requires a \$1000 investment at the beginning of the first year and generates a \$200 revenue at the end of each of the following 5 years.

Project option P3: This opportunity requires a \$2000 investment at the beginning of the first year and generates a variable stream of net revenues at the end of each of the next 5 years as follows: \$1500, \$1000, \$800, \$900, and \$1200.

Project option P4: This opportunity requires a \$5000 investment at the beginning of the first year and generates a variable stream of net revenues at the end of each of the next 5 years as follows: \$1000, \$1500, \$2000, \$3000, and \$4000.

Table 3 Cash Flow and Net Value Calculations of Four Project Options or Proposals, Assuming and MARR of $i = 10\%$

End of Year N	Do-Nothing Option P1	Project Option P2	Project Option P3	Project Option P4
Given Cash Flow				
0	0	-1,000	-2,000	-5,000
1	0	200	1,500	1,000
2	0	200	1,000	1,500
3	0	200	800	2,000
4	0	200	900	3,000
5	0	200	1,200	4,000
Calculations				
Net cash flow ($\sum P$)	0	0	+3,400	+6,500
Net present value at the end of year 5 ($NPV _{N=5}$)	0	-242	+2,153	+3,192
Net present value for revenue to continue to ∞ ($NPV _{n=\infty}$)	0	+1,000	+9,904	+28,030
Average annual ROI ($ROI _{N=5}$)	0	20%	54%	46%
Cost-benefit CB = $ROI_{NPV N=5}$	0	76%	108%	164%
Payback period for MARR=10% $N_{PBP i=10}$	0	8	1.8	3.8
Payback period for MARR = 0% $N_{PBP i=0}$	0	5	1.5	3.3

Note: Given for all four project proposals: (1) a single investment is being made at the beginning of the project life cycle (e.g., at the end of year 0) and (2) the internal rate of return, IRR, or the minimum attractive rate of return, MARR, is 10%.

We can determine the NPV of a single revenue, or stream of future revenues, or costs expected in the future. Two types of presentations are common: (1) present worth and (2) net present value.

Present Worth (PW). This is the single revenue or cost (also called annuity A) that occurs at the end of a period n , subject to the *prevailing interest rate* i . Depending on the management philosophy and enterprise policies, this interest rate can be (i) the *internal rate of return (IRR)* realized by the company on similar investments and (ii) the *minimum attractive rate of return (MARR)* acceptable to company management, or the prevailing discount rate. The present worth is calculated as

$$PW(A|i, n) = PW_n = A \frac{1}{(1+i)^n}$$

For the examples used in this chapter, we consider the internal rate of return, IRR (defined as the average return realized on similar investments), to be the prevailing interest rate.

Net Present Value (NPV). The *net present value* is defined as a series of revenues or costs, A_n , over N periods of time, at a prevailing interest rate i :

$$PW(A|i, n) \sum_{n=1}^N A_n \frac{1}{(1+i)^n} = \sum_{n=1}^N PW_n$$

Three special cases exist for the NPV calculation: (1) for a uniform series of revenues or costs over N periods: $NPV (A_n|i, N) = A[(1+i)^{N-1}]/i(1+i)^N$; (2) for an annuity or interest rate i approaching zero: $NPV = A \times N$; and (3) for the revenue or cost series to continue forever: $NPV = A/i$. Table 3 applies these formulas to the four project alternatives described in Table 2, showing the most favorable 5-year net present value of \$3192 for project option P3.

2.2 Return on Investment Comparison

Perhaps, one of the most popular measures for project evaluation is the *return on investment*, ROI:

$$ROI = \frac{\text{Revenue } (R) - \text{Cost } (C)}{\text{Investment } (I)}$$

ROI calculates the ratio of net revenue over investment. In its simplest form, the stream of cash flow is not discounted. One can look at the revenue on a year-by-year basis, relative to the initial investment. For example, project option 1 in Table 3 would produce a 20% ROI each year, whereas project option 2 would produce a 75% ROI during the first year, 50% during the second year, and so on. In a somewhat more sophisticated way, we can calculate the average ROI per year over a given revenue cycle as shown in Table 3:

$$\overline{ROI}(A_n I_n | N) = \left[\frac{\sum_{n=1}^N (\text{Revenue } R)_n - (\text{Cost } C)_n}{(\text{Investment } I)_n} \right]$$

We can then compare the average ROI to the MARR. Given a MARR of 10% for our project environment, all three project options P1, P2, and P3 compare favorable, with project P3 yielding the highest average return on investment of 54%. Although this is a popular measure, it does not permit a meaningful comparative analysis of alternative projects with fluctuating costs and revenues. Furthermore, it does not consider the time value of money.

2.3 Cost–Benefit

Alternatively, we can calculate the net present value of the total ROI over the project life cycle. This measure, known as *cost–benefit (CB)*, is calculated as the present value stream of net revenues divided by the present value stream of investments. It is an effective measure for comparing project alternatives with fluctuating cash flows:

$$CB = ROI_{NPV}(A_n, I_n | i, N) = \left[\sum_{n=1}^N NPV(A_n | i, N) \right] \div \left[\sum_{n=1}^N NPV(I_n | i, N) \right]$$

In our example of four project options (Table 3), project proposal P4 produces the highest cost–benefit of 164% under the given assumption of $i = MARR = 10\%$.

2.4 Payback Period Comparison

Another popular figure of merit for comparing project alternatives is the payback period (PBP). It indicates the time period of net revenues required to return the capital investment made on the project. For simplicity, undiscounted cash flows are often used to calculate a quick figure for comparison, which is quite meaningful if we deal with an initial investment and a steady stream of net revenue. However, for fluctuating revenue and/or cost streams, the net present value must be calculated for each period individually and cumulatively added up to the “break-even point” in time, N_{PBP} , when the net present value of revenue equals the investment. Mathematically, N_{PBP} occurs when

$$\sum_{N=1}^N NPV(A_n | i) \geq \sum_{n=1}^N NPV(I_n, J_i)$$

In our example of four project options (Table 3), project proposal P3 produces the shortest, most favorable payback period of 1.8 years under the given assumption of $i = MARR = 10\%$.

2.5 Pacifico and Sobelman Project Ratings

The previously discussed methods of evaluating projects rely heavily on the assumption that technical and commercial success is assured, and all costs and revenues are predictable. Because these assumptions do not always hold, many companies have developed their own special procedures and formulas for comparing project alternatives. Two examples illustrate this special category of project evaluation metrics.

The Project Rating (PR) Factor. This measure was originally developed by Carl Pacifico for assessing chemical products and predicting commercial success:

$$PR = \frac{pT \times pC \times R}{TC}$$

Pacifico’s formula is in essence an ROI calculation adjusted for risk. It includes probability of technical success ($0.1 < pT < 1.0$), probability of commercial success ($0.1 < pC < 1.0$), total net revenue over project life cycle (R), and total capital investment for product development, manufacturing setup, marketing, and related overheads (TC).

Product Development Figure of Merit. The formula developed by Sobelman

$$z = (P \times T_{LC}) - (C \times T_D)$$

represents a modified cost–benefit measure that takes into account both the development time and the commercial life cycle of the product. It also includes average profit per year (P), estimated product life cycle (T_{LC}), average development cost per year (C), and years of development (T_D).

Table 4 Comparison of Quantitative and Qualitative Approaches to Project Evaluation

Quantitative Methods	Qualitative Methods
Benefits	Benefits
Clear and simple comparison, ranking, selection	Search for meaningful evaluation metrics
Repeatable process	Broad-based organizational involvement
Encourages data gathering and measurability benchmarking opportunities	Understanding of problems, benefits, opportunities
Programmable	Problem solving as part of selection process
Useful input to sensitivity analysis and simulation	Broadly distributed knowledge base
Connectable to many analytical and statistical models	Multiple solutions and alternatives
	Multifunctional involvement leading to buyin and risk sharing
Limitations	Limitations
Many success factors are not quantifiable	Complex, time-consuming process
Probabilities and weights may change	Biases introduced via organizational power
True measures do not exist	Difficult to procedurize or repeat
Analyses and conclusions are often misleading	Conflict and disagreement over decision/outcome
Masking of hidden problems and opportunities	Does not fit conventional decision processes
Stifle innovative decision making	Intuition and emotion may obscure facts
Lack people involvement, buy-in, commitment	Used for justifying “wants”
Ineffective in dealing with multifunctional issues, nonlinearities, and dynamic situations	Lead to more fact finding than decision making
May mask hidden costs and benefits	Temptation for unnecessary expansion of fact finding
Temptation for acting too quickly and prematurely	Process requires effective managerial leadership

2.6 Going Beyond Simple Formulas

While quantitative methods of project evaluation have the benefit of producing relatively quickly a measure of merit for simple comparison and ranking, they also have many limitations, as summarized in Table 4. Yet, in spite of the limitations inherent to quantitative evaluation and the increased use of qualitative approaches, virtually every organization supports its project selections with some form of quantitative measures—most popular ROI, cost–benefit, and payback period. However, driven by the growing complexity of the business environment, managers are getting increasingly concerned about these limitations and explore alternatives. They often augment quantitative methods with additional measures for determining the long-range cost–benefits of a project proposal to the enterprise. Many of these contemporary decision-making methods rely to a large degree on qualitative, judgmental decision making. These data gathering methods cast a wide net and consider a broad spectrum of factors that are often difficult to describe or quantify but are effective in gaining strategic perspective and a more comprehensive picture on potential benefits, risks, and challenges of the proposed project.

3 QUALITATIVE APPROACHES

While quantitative methods provide an important tool set for project evaluation and selection, there is also a growing sense of frustration, especially among managers of complex and technologically advanced undertakings, that reliance on strictly quantitative methods does not always produce the most useful or reliable inputs for decision making, nor are all methods

equally suited for all situations.^{16,17} Therefore, it is not surprising that for project evaluations involving complex sets of business criteria, narrowly focused quantitative methods are often supplemented with broad-scanning, intuitive processes and collective, multifunctional decision making such as Delphi, nominal group technology, brainstorming, focus groups, sensitivity analysis, and benchmarking. Each of these techniques can either be used by itself to determine the “best, most successful, or most valuable” option or integrated into a comprehensive analytical framework for collective multifunctional decision making, which is being discussed in the following section.

3.1 Collective Multifunctional Evaluations

This process relies on subject experts from various functional areas for collectively defining and evaluating broad project success criteria, employing both quantitative and qualitative methods.¹⁷ The first step is to define the specific organizational areas critical to project success and to assign expert evaluators. For a typical product development project, these organizations may include R&D, engineering, testing, manufacturing, marketing, product assurance, and customer/field services. These function experts should be given the time necessary for the evaluation. They also should have the commitment from senior management for full organizational support. Ideally, these evaluators should be members of the core team ultimately responsible for project implementation.

Evaluation Factors. Early in the evaluation process, the team defines the factors that appear critical to the ultimate success of the projects under evaluation and arranges them into a list, which includes both quantitative and qualitative factors. A mutually acceptable scale must be worked out for scoring the evaluation criteria. Studies of collective multifunctional assessment practices show that simple scales are most effective for leading to actionable team decisions. The four most popular and robust scales for judging situational outcomes are as follows:

1. *Ten-Point Judgment Scale.* From +5 (most favorable) to –5 (most unfavorable)
2. *Three-Point Judgment Scale.* +1 (favorable), 0 (neutral or cannot judge), and –1 (unfavorable)
3. *Five-Point Judgment Scale.* A (highly favorable), B (favorable), C (marginally favorable), D (most likely unfavorable), and F (definitely unfavorable)
4. *Five-Point Likert Scale.* 1 (strongly agree), 2 (agree), 3 (neutral), 4 (disagree), and 5 (strongly disagree)

Weighing of criteria is not recommended for most applications as it complicates and often distorts the collective evaluation.

The Evaluation Process. Evaluators first assess and then score all of the success factors they feel qualified to judge. Then collective discussions follow. Initial discussions of project alternatives—their markets, business opportunities, and technologies involved—are usually beneficial but not necessary for the first round of the evaluation process. The objective of this first round of expert judgments is to get calibrated on the opportunities and challenges presented. Further, each evaluator has the opportunity to recommend (1) actions that could improve the quality and accuracy of the project evaluation, (2) additional data needed, and (3) suggestions for increasing project success. Before meeting at the next group session, agreed-on action items and activities for improving the decision process should be completed. The evaluation process is enhanced with each iteration by producing more accurate, refined and comprehensive data. Typically, between three and five iterations are required before a go/no-go decision can be reached for a given project.

4 RECOMMENDATIONS

Effective evaluation and selection of project opportunities involves many variables of the organizational and technological environment, reaching often far beyond cost and revenue measures. While economic models provide an important dimension of the project selection process, most situations are too complex to use simple quantitative methods as the sole basis for decision making. Many of today's project evaluation procedures include a broad spectrum of variables and rely on a combination of rational and intuitive processes for defining the value of a new project venture to the enterprise. The better an organization understands its business processes, markets, customers, and technologies, the better it will be able to evaluate the value, risks, and challenges of a new project venture. Further, manageability of the evaluation process is critical to its results, especially in complex situations. The process must have a certain degree of structure, discipline, and measurability to be conducive to the intricate multivariable analysis. One method of achieving structure and manageability calls for grouping the evaluation variables into four categories: (1) consistency and strength of the project with the business mission, strategy, and plan; (2) multifunctional ability to produce the project deliverables and objectives, including technical, cost, and time factors; (3) success in the customer environment; and (4) economics, including profitability. Modern phase management, such as stage-gate processes provide managers with the tools for organizing and conducting project evaluations in a systematic way. The following section summarizes suggestions that can help managers in effectively evaluating and selecting projects toward successful implementation:

Seek-Out Relevant Information. Meaningful project evaluations require relevant quality information. The four sets of variables, related to the strategy, results, customer and economics, as identified above, can provide a framework for establishing the proper metrics and detailed data gathering.

Ensure Competence and Relevancy. Ensure that the right people become involved in the data collection and judgmental processes.

Take Top-Down Look first, Detail Comes Later. Detail is less important than information relevancy and evaluator expertise. Do not get hung-up on missing data during the early phases of the project evaluation. Evaluation processes should be iterative. It does not make sense to spend a lot of time and resources on gathering perfect data, to justify a "no-go" decision.

Select and Match the Right People. Whether the project evaluation consists of a simple economic analysis or a complex multifunctional assessment, competent people from functions critical to the overall success of the project should be involved.

Define Success Criteria. Whether deciding on a single project or choosing among alternatives, evaluation criteria must be defined. They can be quantitative, such as ROI, or qualitative, such as the chances of winning a contract. In either case, these evaluation criteria should cover the true spectrum of factors affecting success and failure of the project(s). The success criteria should be identified by seasoned enterprise personnel. In addition, people from outside of the company, such as vendors, subcontractors, and customers, are often included in this expert group and critical to the development of meaningful success criteria.

Strictly Quantitative Criteria can be Misleading. Be aware of evaluation procedures based on quantitative criteria only (ROI, cost, market share, MARR, etc.). The input data used to calculate these criteria are likely based on rough estimates and are often unreliable. Furthermore, a reliance on strictly quantitative data, considers only a narrow spectrum of factors affecting

project success or failure, thus ignoring many other important factors, especially those that influence project success in a dynamic or nonlinear way, typical for many complex technologically sophisticated undertakings. Evaluations based on predominately quantitative criteria should at least be augmented with some expert judgment as a “sanity check.”

Condense Criteria List. Combine evaluation criteria, especially among the judgmental categories, to keep the list manageable. As a goal, try to stay within 12 criteria for each category.

Gain Broad Perspective. The inputs to the project selection process should include the broadest possible spectrum of data from the business environment that affect success, failure, and limitations of the new project opportunity. Assumptions should be carefully examined.

Communicate Across the Enterprise. Facilitate communications among evaluators and functional support groups. Define the process for organizing the team and conducting the evaluation and selection process.

Ensure Cross-Functional Representation and Cooperation. People on the evaluation team must share a strategic vision across organizational lines. They also must have the desire to support the project if selected for implementation. The purpose, goals, objectives, and relationships of the project to the business mission should be clear to all parties involved in the evaluation/selection process.

Do not Lose the Big Picture. As discussions go into detail during the evaluation, the team should maintain a broad perspective. Two global judgment factors can help to focus on the big picture of project success: (1) overall cost–benefit perspective and (2) overall risk of failure assessment. These factors can be recorded on a 10-point scale, –5 to +5. This also leads to an effective two-dimensional graphic display for comparing competing project proposals.

Do Your Homework between Iterations. Project evaluations are usually conducted progressively in iterative cycles. Therefore, the need for more information, clarification, and further analysis surfaces between each cycle. Necessary action items should be properly assigned and followed up to enhance the evaluation quality with each consecutive iteration.

Take a Project-Oriented Approach. Plan, organize, and manage your project evaluation/selection process as a project. Proposal evaluation and selection processes require valuable resources that must be justified and carefully managed.

Resource Availability and Timing. Do not forget to include in your selection criteria the availability and timing of resources. Many otherwise successful projects fail because they cannot be completed within a required time period.

Use Red Team Reviews. Set up a special review team of senior personnel. This is especially useful for large and complex projects with major impact on overall business performance. This review team examines the decision parameters, qualitative measures, and assumption used in the evaluation process. Limitations, biases, and misinterpretations that may otherwise remain hidden can often be identified and resolved.

Stimulate Creativity and Candor. Senior management should foster an innovative risk-shared ambience for the evaluation team. Especially, the evaluation of complex project situations involves intricate sets of variables. Criteria for success and failure are linked among many subsystems, such as organization, technology and business, associated with a great deal of

risks, and uncertainty. Innovative approaches are required to evaluate the true potential of success for these projects. Risk sharing by senior management, recognition, visibility and a favorable image in terms of high priority, interesting work, and importance of the project to the organization have been found strong drivers toward attracting and holding quality people on the evaluation team and toward gaining their active and innovative participation in the process.

Manage and Lead. The evaluation team should be chaired by someone who has the trust, respect, and leadership credibility with the team members. Senior management can positively influence the work environment and the process by providing guidelines, charters, visibility, resources, and active support to the project evaluation team.

In summary, effective project evaluation and selection requires a broad-scanning process across all segments of the enterprise and its environment to deal with the risks, uncertainties, ambiguities, and imperfections of data available for assessing the value of a new project venture relative to other opportunities. No single set of broad guidelines exist that guarantees the selection of successful projects. However, the process is not random! A better understanding of the organizational dynamics that affects project performance and the factors that drive cost, revenue, and other benefits can help in gaining a better, more meaningful insight into the future value of a prospective new project. Seeking out both quantitative and qualitative measures incorporated into a combined rational–judgmental evaluation process often yield the most reliable predictor of future project value and desirability. As equally important, the process requires managerial leadership and skills in planning, organizing, and communicating. Above all, the leader of the project evaluation team must be a social architect, who can unify the multifunctional process and its people. The leader must be able to foster an environment professionally stimulating and conducive to risk sharing. It also must be effectively linked to the functional support groups needed for project implementation. Finally, organizational strategy must be aligned and integrated with the evaluation/selection process, early and throughout its evaluation cycle. Senior management has an important role in unifying the evaluation team behind the mission objectives and in facilitating the linkages to the stakeholders and ultimate user community. Senior management should further help in providing overall leadership, and in building mutual trust, respect and credibility among the members of the proposal evaluation team, all critical drivers toward a strong partnership of all team members and the basis for an effective enterprise-wide decision-making system. Taken together, this is the environment conducive to cross-functional communication, cooperation, and integration of the intricate variables needed for effective engineering project evaluation and selection.

5 TERMS

Cross Functional: Actions that span organizational boundaries.

Phase Management: Projects are broken into natural implementation phases, such as development, production, and marketing, as a basis for project planning, integration, and control. Phase management also provides the framework for *concurrent engineering* and *stage-gate processes*.

Project Success: A comprehensive measure, defined in both quantitative and qualitative terms, that includes economic, market, and strategic objectives.

Stage-Gate Process: Framework originally developed by R. Cooper and S. Edgett for executing projects within predefined stages (See also *phase management*) with measurable deliverables (at gates) at the end of each stage. These gates also provide the review metrics for ensuring successful transition and integration of the project into the next stage.

Weighing of Criteria: A multiplier associated with specific evaluation criteria.

6 VARIABLES AND ABBREVIATIONS

Annuity (A) is the present worth of a revenue or cost at the end of a period n .

Cost–benefit (CB), net present value of all ROIs in dollars.

Prevailing interest rate (i)

Investment (I)

Internal rate of return (IRR), the average return on investment realized by a firm on its investment capital MARR (minimum attractive rate of return) on new investments acceptable to an organization.

Net present value (NPV) of a stream of future revenues or costs.

Payback period (PBP), the time period needed to recover the original investment.

Project rating (PR) factor, a measure developed by Carlo Pacifico for predicting project success.

Present worth (PW) (also called annuity), the present value of a revenue or cost at the end of a period n .

Return on investment (ROI)

Project rating factor (z), a measure developed by Sobelman for predicting project success.

REFERENCES

1. G. Prabhakar, "What Is Project Success: A Literature Review," *Int. J. Bus. Manage.*, **3**(9), 3–10, 2009.
2. T. Raz, A. Shenhar, and D. Dvir, "Risk Management, Project Success and Technological Uncertainty," *R&D Manage.*, **32**(2), 101–109, 2002.
3. A. Shenhar, D. Dvir, L. Ofer, and A. Maltz, "Project Success: A Multidimensional Strategic Concept," *LongRange Planning*, **34**(6), 699–725, 2002.
4. H. Thamhain, and T. Skelton, "Success Factors for Effective R&D Risk Management," *Int. J. Tech. Intell. Planning (IJTIP)*, **3**(4), 376–386, 2007.
5. S. Cicmil, T. Williams, J. Thomas, and D. Hodgson, "Rethinking Project Management: Researching the Actuality of Projects," *Int. J. Proj. Manage.*, **24**(8), 675–686, 2006.
6. W. F. Lemon, J. Bowitz, J. Burn, and R. Hackney, "Information Systems Project Failure: A Comparative Study of Two Countries," *J. Global Info. Manage.*, **10**(2), 28–39, 2002.
7. K. El Eman, and A. Koru, "A Replicated Survey of It Software Project Failures," *Software (IEEE)*, **25**(5), 84–90, 2008.
8. Shenhar and Dvir, *Reinventing Project Management: The Diamond Approach to Successful Growth and Innovation*. Harvard Business School Press, Boston, MA, 2007.
9. L. Bstieler, "The Moderating Effects of Environmental Uncertainty on New Product Development and Time Efficiency," *J. Product Innovation Manage.*, **22**(3), 267–284, 2005.
10. S. Cicmil, and D. Marshall, "Insights into Collaboration at the Project Level: Complexity, Social Interaction and Procurement Mechanisms," *Building Res. Info.*, **125**(6), 627–668, 2005.
11. D. S. Remer, S. B. Stokdyk, and M. Van Driel, "Survey of Project Evaluation Techniques Currently Used in Industry," *Int. J. Product. Econ.*, **32**(1), 103–115, 1993.
12. S. Kavadias, and C. H. Loch, *Project Selection under Uncertainty: Dynamically Allocating Resources to Maximize Value*. Kluwer Academic, Norwood, MA, 2004.
13. R. Khorramshahgol, H. Azani, and Y. Gousty, "An Integrated Approach to Project Evaluation and Selection," *IEEE Trans. Eng. Manage.*, **35**(4), 265–270, 1998.
14. H. Thamhain, "Project Evaluation and Selection," in *Management of Technology*, Chapter 8, Wiley, Hoboken, NJ, 2005.
15. S. Mantel, J. Meredith, S. Shafer, and M. Sutton, "Selecting Projects to Meet Organizational Objectives," in *Project Management Practice*, Chapter 1.5, Wiley, Hoboken, NJ, 2011, pp. 10–22.
16. R. B. Kulkarni, D. Miller, R. M. Ingram, C.-W. Wong, and J. Lorenz, "Need-Based Project Prioritization: Alternative to Cost-Benefit Analysis," *J. Eng. Transport.*, **130** (2), 150–158, 2004.
17. P. D. Kumar, "Integrated Project Evaluation and Selection Using Multiple-Attribute Decision-Making Technique," *Int. J. Product. Econ.*, **103**(1), 87, 2006.

CHAPTER 21

LEAN MANAGEMENT

Eric H. Stapp and Cynthia M. Sabelhaus
Raytheon Missile Systems Company
Tucson, Arizona

1 INTRODUCTION	617	4.3 Poka-Yoke	623
2 BRIEF HISTORY OF LEAN	617	4.4 Total Productive Maintenance	623
2.1 Toyota Method	618	4.5 Just-in-Time	623
3 BASIC LEAN PHILOSOPHY AND DEPLOYMENT	618	4.6 Kanban	623
3.1 Value Stream	618	4.7 Takt Time	624
3.2 Muda	619	4.8 Gemba	624
3.3 Kaizen	620	5 TRADITIONAL QUALITY CONTROL TOOLS	624
3.4 Plan–Do–Check–Act Approach	620	6 SEVEN MANAGEMENT TOOLS	629
3.5 Kaizen Event or Blitz	621	7 LEADERSHIP AND LEAN	633
4 LEAN TOOLS	621	REFERENCES	634
4.1 5S	621		
4.2 Visual Factory	622		

1 INTRODUCTION

Lean management is a philosophy that encompasses all activities from manufacturing to leadership to supply chain interaction. This philosophy includes a number of tools and methods for its implementation and has been responsible for saving costs and improving quality and profitability in U.S. organizations. Its portability to any endeavor makes “lean” a valuable toolset for the mechanical engineer.

2 BRIEF HISTORY OF LEAN

Toyota Motor Company has a long and interesting history. It began in the Aichi prefecture of Japan 80 years ago, and it was profitable every year from 1950 until 2008, when the combined effects of a worldwide recession, devastating tidal wave, and large auto recall caused the company to experience losses. Toyota was able to turn even these events around and was profitable again in 2010. Much of this long-sustained profitability is attributed to the Toyota leadership model and lean manufacturing.

The Toyoda Automatic Loom Works began operations in 1896. Its founder, Sakichi Toyoda, was looking for a way to help his mother and grandmother increase their cloth-making efforts that supplemented the family’s income. By 1924, Sakichi had incorporated foot pedals to move the shuttle of threads back and forth, eliminating much of the handwork. He later incorporated steam engine technology to automate the wooden looms. Platt Brothers of England bought the loom company, providing the funds that started the Toyota Motor Company.

Kiichiro Toyoda, son of Sakichi, and Eiji Toyoda, a nephew, started the Toyota automotive company in 1930. It was a time of constrained resources in Japan and Kiichiro was forced to examine all inputs and movements in his auto company just to stay in business. This laid the foundation for Toyota Production Systems (TPS) that was later dubbed Lean Manufacturing. Kiichiro traveled to Europe and the United States to examine their engine designs and made many innovative changes to create an engine that would work with the available parts and raw materials in Japan. During the Pacific War (World War II), Toyota Motor Company was engaged so heavily in building trucks for military use that its auto manufacturing almost disappeared.

After World War II, Toyota examined its processes as part of the nationwide rebuilding effort. Eiji Toyoda spent three months at the Ford Motor Company in Detroit. At that time, Ford was engaged in large-scale assembly line processes, producing 8000 cars a day, which was not appropriate for Toyota Motors, with an annual production of 2500. After his return to Japan, Eiji along with Toyota's production manager, Taiichi Ohno, realized that Ford's mass production would not work at Toyota. The two developed a new means of production that included engineering, manufacture, supply, assembly, and workforce management.¹

2.1 Toyota Method

The results of Eiji's and Ohno's efforts eventually became known as the TPS, and as the company expanded its markets to offer its highly reliable products to the United States, they were visited by countless American companies wanting to learn the TPS methods. In 1988, a team of MIT researchers coined the term *lean manufacturing*. It is synonymous with TPS.

3 BASIC LEAN PHILOSOPHY AND DEPLOYMENT

Basic to the TPS is the elimination of waste in the value stream. Waste encompasses not only non-value-added activities such as rework but also travel, waiting, storing excess inventory, and failing to take advantage of the full capabilities and potential of each employee. To identify waste, it is first essential to understand the value stream.²

3.1 Value Stream

The value stream is composed of all things that go into a product or service, whether those things add value or not. To define the value stream for a product or service, there must be clear understanding of all components involved in producing the product, including materials, labor, and information.

To understand the value stream, the customers' needs and wants in the products or services must be understood. This is not always easy. For example, if a customer arrives at the grocery store to buy a banana, the store manager might think that the customer has received the value he or she expected if the banana is available, of good quality, and priced reasonably. On the other hand, these elements meet only part of the customer's needs. The customer may have additional needs that are not readily apparent but when they are not met will discourage him or her from returning to the store for another banana or anything else. If the store is found to be dirty, the staff unfriendly, and the lines at the checkout counter long, the customer's needs will not have been satisfied. To understand the value stream for delivering the banana to a customer, all parts of the process must be examined from the customer's point of view.

To ensure a complete understanding of what the customer values in each product or service, surveys, focus groups, and other techniques are used. It is also necessary to understand that customer values change with time and circumstances. One clear example is the type of automobile a customer seeks and its correlation with the price and availability of gasoline.

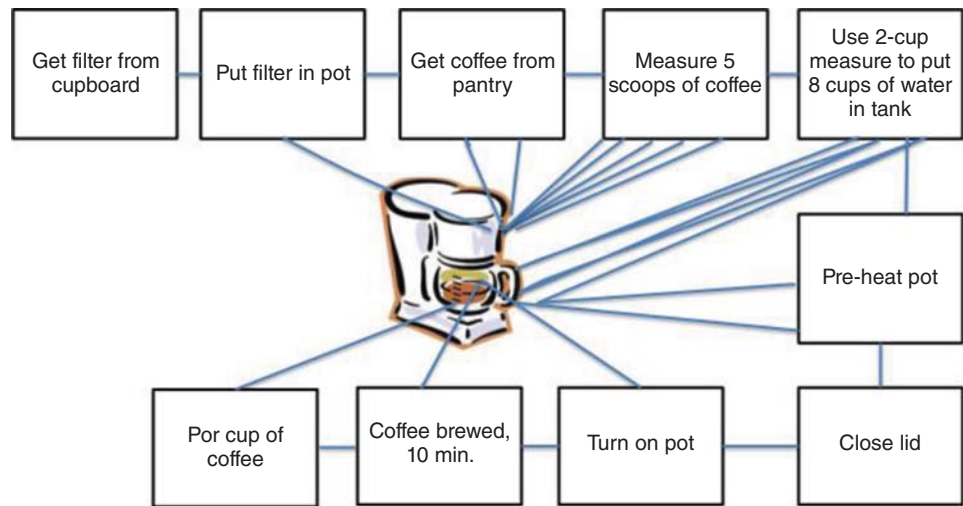


Figure 1 Value stream for making a pot of coffee. A spaghetti chart is superimposed, showing the movement for each step.

When there is a threat to the supply of gasoline, customers quickly change the value they seek in a new car to include gas mileage or alternative fuels.

Once the value stream is understood, it can be mapped, including each step in the processes that results in delivery of the product or service to the customer. This value stream map can serve as the first step in leaning out the process and eliminating waste, also known as *muda*. In Fig. 1 a simple value stream map shows the steps in making a pot of coffee, with the end “product” being a cup of coffee from the pot. In order to see how the process is carried out, a spaghetti chart is superimposed, using lines to show the movement between the coffee pot and each step in the process. It is easy to see that filling a two-cup measure four times results in eight trips between pot and faucet. There is also wait time as the cup is filled each time. Wait time is wasted time.

3.2 Muda

Muda is the Japanese word for waste, and central to TPS is eliminating waste. When the value stream has been defined, it becomes clear that many of the activities required to produce the product or service do not contribute to its ultimate value. Some activities are necessary although they are still waste when put in the context of contributing to the end value. For example, manufactured items are likely to be inspected at some point in the manufacturing process to ensure they meet requirements. Such inspections are *muda* because they do not contribute to the product’s creation. However, the inspection may not be eliminated unless all potential errors have been eliminated, and that may be impossible.

Muda was further divided into seven types by Taiichi Ohno in 1988.² Ohno’s types of *muda* are:

1. Defects
2. Overproduction
3. Inventories (in process or finished goods)
4. Unnecessary processing

5. Unnecessary movement of people
6. Unnecessary movement of goods
7. Waiting

In 1996 Womack and Jones³ added an eighth type of *muda*:

8. Designing goods and services that do not meet customers' needs

Lean Master Shigeo Shingo defined another type of waste:

9. *Underutilization of People*. This differs from waiting in that it encompasses the failure to allow people to fully use their talents, contribute their ideas, or use their full energy in their work environment.

In addition to the seven or eight or nine types of *muda*, there are other types of waste:

Mura (unevenness) is waste due to variation in quality, cost, or delivery. *Mura* happens when quality cannot be predicted. This is the waste associated with additional testing or inspecting and the waste produced when something is returned or must be reworked because it was not built to specifications in the first place. *Mura* is reduced through standardization of processes.

Muri (overdoing) is waste due to unnecessary overburdening of people, equipment, or systems, often in an attempt to exceed capacity. The result is workplace injury, broken machinery, and ultimately the waste of waiting while some part of the production system is repaired and brought back on line.

3.3 Kaizen

Kaizen comes from two Japanese words: *kai* (change) and *zen* (to see, or to gain wisdom from doing). The two words taken together have been translated as “change for the better or continuous improvement.” Kaizen is more an underlying philosophy than an improvement tool, although deployed properly, it uses many improvement tools and methods. It is based on the philosophy that each person in an organization can improve their work processes and that each person will approach each day as an opportunity to make improvements.

Kaizen differs from Six Sigma and other improvement methodologies in that it involves everyone from the CEO to someone in an entry-level position. Except for some training, Kaizen happens within the job or work unit and is not a separate activity. Changes are incremental rather than monumental and often require few resources. For example, in the coffee value stream, a Kaizen improvement might be as simple as moving the coffee closer to the maker or using a larger scoop to reduce the number of scoops or even creating premeasured coffee packets.

In addition to creating continual, incremental improvements in every facet of a business, Kaizen also values process consistency and calls for the establishment of policies and rules to maintain the performance levels set within the current process. When changes are made, these policies must be updated so that everyone consistently follows the same steps, thereby achieving a consistent result.

In many cases, Kaizen is deployed following widespread training for all workers. In some of Japan's companies known for the quality of their products such as Toyota and Canon, employees are expected to generate one hundred improvement suggestions a year. Setting improvement goals reinforces the practice of continuous improvement in the culture.⁴

3.4 Plan–Do–Check–Act Approach

Although the plan–do–check–act (PDCA) cycle was not born in Japan, it is sometimes called the lean operating framework. PDCA was first defined in the United States in the 1930s by

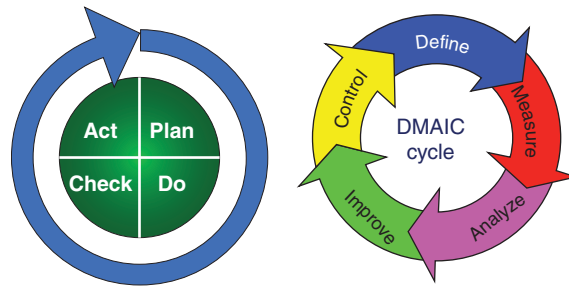


Figure 2 PDCA and DMAIC improvement cycles.

Walter Shewhart at Bell Telephone. W. Edwards Deming used a similar approach in the 1950s that he called plan–do–study–act (PDSA), and many practitioners of improvement methods have added an O for “observe” in front of the four-step process. Six Sigma uses a similar model of define–measure–analyze–improve–control (DMAIC). Regardless of the labels, these methods are all linked to the scientific discipline of forming a hypothesis, running an experiment, and evaluating the results (see Fig. 2).

The PDCA cycle is used for each lean improvement. An improvement suggestion is made and approved (plan), the improvement is made (do), the results of the improvement are measured (check), and if there is evidence of a sustained improvement in the outcome, the new method is incorporated into the process (act).

3.5 Kaizen Event or Blitz

While Kaizen is a philosophy of continuous improvement, Kaizen events or blitzes are discrete activities that bring people together to quickly improve the value stream. The Kaizen event is also called a workshop or blitz and consists of focusing everyone on a particular portion of the value stream in order to eliminate *muda*. The events are typically of short duration—five or fewer work days—with improvement activities running the full PDCA cycle during the span of the workshop. In order to realize the quickest gains from the Kaizen event, the scope of each project must be carefully controlled. Scope creep is often the root cause of failed projects. Participants in the event may require a full day of training on improvement techniques before the event can begin to address improvements.

Table 1 provides a sample agenda for a Kaizen event.

4 LEAN TOOLS

4.1 5S

5S is a workplace organization method. The name stems from five Japanese words: *seiri*, *seiton*, *seiso*, *seiketsu*, and *shitsuke*. These translate loosely as:

1. Sorting—determining which items in the work area should stay and which can be eliminated or stored elsewhere.
2. Straightening or setting in order—once items are determined to be required in the work area, they are placed in a way to facilitate use and consistent storage.
3. Sweeping or shining—after all the items needed at the workstation are put away, the area is thoroughly cleaned.

Table 1 Kaizen Event Agenda

Day	Agenda	Tools
1	Training	Team building exercises Tool training Value stream Kaizen philosophy 5S 7 Wastes Measurement systems Takt time Improvement tools
2	Review value stream map Analyze current process Identify area to improve Brainstorm improvements Design new work methods	Spaghetti chart Pre-Kaizen metrics Quality function deployment 5 whys; fishbone diagram
3	Implement new work methods	Measure results
4	Observe new work methods Refine work methods Measure results	Documented work standards
5	Finalize work standards Celebrate results	Visual controls Post-Kaizen metrics

4. Standardizing—all workstations for a particular job are arranged identically. All employees doing the same job should be able to work in any station with the same tools that are in the same location in every station. Everyone should know exactly what his or her responsibilities are for adhering to the first 3 S's.
5. Sustaining the practice—once workstations are arranged in the most efficient manner, this arrangement must be maintained.

Additional S's are sometimes added, most commonly safety, security, and satisfaction.

A 5S activity will require stopping work for a defined period of time to allow for the sorting, storing, shining, and other steps. The outcome from the activity can yield great improvements in productivity and quality, but sustainment is often an issue. A year after a 5S blitz has taken place, there may be no evidence that the effort was made. Tools may be scattered around the workplace or even lost. Items that are not being used in the processes in the area are again stacked in corners and on shelves. Clearly, without constant vigilance, the labor used in the 5S activity could have been better spent on making products, and the gains from the engagement are never fully realized.⁵

4.2 Visual Factory

The visual factory is self-explanatory. As much as possible, item storage, traffic flows, and work instructions are visual. Signs and floor paint are used to mark traffic patterns. Tools are kept in foam cutouts so that a missing tool is quickly noticed and found. Defects and inventory issues are identified by a light or buzzer or both to alert the work unit that an event has occurred and help is required. Storage in a visual factory is often color coded to improve retrieval and prevent items from being placed in the wrong spot.

4.3 Poka-Yoke

Poka-yoke is the act of error proofing each step in the workflow. A clear example of poka-yoke is the way most computers and peripherals are designed so that each item to be plugged into the computer only fits in one place. The work of assembling the computer and peripherals has been error proofed to a large extent.

When defects are found in products, it is not uncommon for management to attribute the cause to “operator error,” and the corresponding corrective action is to retrain the operator. With poka-yoke, the individual is not blamed for making the error. It is assumed that if one person can make the mistake that caused the defect, then any person doing that job could make the same mistake. To correct the problem, a way must be found to prevent anyone from making the error.

One simple example of poka-yoke resulted from bank employees sending statements to the wrong customers. The process called for preaddressed envelopes into which an employee inserted a bank statement. In 1% of the cases, the wrong statement went into the envelope, thus delivering one customer’s private financial information to another customer. By changing to window envelopes, the only address on the envelope was the address on the statement. The process was error proofed so that a customer could not receive another customer’s statement.

4.4 Total Productive Maintenance

Total productive maintenance (TPM) is the process of predicting maintenance needs and then taking appropriate action before the equipment breaks down and affects the workflow. TPM schedules are developed based on historical data, manufacturers’ specifications, and unusual environmental conditions. By using TPM, maintenance can be scheduled to prevent machine down time and level load the workload for maintenance personnel, which may lead to reduced head count requirements.

4.5 Just-in-Time

Just-in-time (JIT) manufacturing is concerned with setting up the entire value stream to eliminate excess production and inventory. The savings from a JIT system are obvious, but realizing them requires careful planning and discipline. For work-in-process inventory, it often requires agreements with suppliers who must be willing to ship small batches of its stock on a more frequent basis. Ideally, the parts will arrive at the facility only a day or two before they are used in the assembly process.

JIT also affects the workflow in the factory. Parts are not delivered to the work area until they are needed and are often placed in clear sight on a shelf as part of the visual factory. When assembly A is moved to workstation 2, workstation 1 should be placing another assembly A in its place on the shelf. If a problem is encountered at workstation 2, the operator at workstation 1 will have to wait until it is resolved before moving its completed work to the staging area or shelf for workstation 2.

4.6 Kanban

Kanban, which means “signboard” in Japanese, is a scheduling system for just-in-time and lean manufacturing. In the late 1940s, Toyota noticed that the shelf stocking methods in a grocery store might have value in a manufacturing environment. If each process served as a customer for all the processes that came before, and if each process took only what it needed from the “store” shelves because the future supply was assured, work could be scheduled based on the pace of each process and its need for the products from preceding processes.

Kanban has six rules:

1. Do not send defective products to the subsequent process.
2. The subsequent process comes to withdraw only what is needed.
3. Produce only the exact quantity withdrawn by the subsequent process.
4. Level the production.
5. Kanban is a means to fine tune.
6. Stabilize and rationalize the process.

Many manufacturers have implemented electronic kanban or e-kanban systems. E-kanban systems can be integrated into enterprise resource planning (ERP) systems, enabling real-time demand signaling across the supply chain and improved visibility. Data pulled from e-kanban systems can be used to optimize inventory levels by better tracking supplier lead and replenishment times.⁶

4.7 Takt Time

The easiest way to avoid overproduction is to base production scheduling on customer demand. Takt time comes from the German word *taktzeit*, which means “cycle time”; however, takt time is always somewhat longer than cycle time to complete the production steps. Takt time includes preparation, packaging, setup, etc. Takt time allows an enterprise to determine whether its available time is appropriate for production of the customer’s desired quantity of the items being produced. To calculate takt time,

$$\text{Takt time} = \text{time available} \div \text{units required}$$

For example, if a customer order required 1000 units to be delivered by the end of the week and there were 40 h (2400 min) in the work week, the time to produce each unit would be

$$\text{Takt Time} = \frac{2400}{1000} = 2.4 \text{ minutes per item}$$

4.8 Gemba

The place where the “real” business occurs, a *gemba walk* is not far different from the “management by walking around” first used by Hewlett-Packard in the 1970s and later described by Tom Peters and Robert H. Waterman in their best-selling *In Search of Excellence*.⁷ Gemba walks allow managers to see the current state and visualize improvements. The walk should not be rushed. A manager doing a gemba walk may wander without agenda through a work area, talking to employees and gaining a better understanding of the work flow, issues, and possible improvements. Gemba walks can also be focused on a particular type of discipline, such as safety, security, or productivity. As an added benefit in performing gemba walks, the manager reinforces the importance of lean practices and continuous improvement just by being present and listening to employees.⁸

Genchi genbutsu is similar to gemba. The literal translation is “go and see.” *Genchi genbutsu* is the act of following up a problem by going to the site of that problem and talking to the people involved. Communication and understanding are enhanced when everyone involved has the same picture of actual events.

5 TRADITIONAL QUALITY CONTROL TOOLS

As the total quality management movement took hold in the 1970s and 1980s, a group of graphical quality control techniques arose as primary data collection and process improvement tools.

Originally, a group of seven tools were identified, but the list grew over time to include additions and variations on the original seven. Some of these tools are quite simple to learn and use and are appropriate for use by those on the factory floor. Others require additional knowledge and experience and are better handled by experienced practitioners.⁹

The original seven tools include:

1. *Cause-and-Effect Diagram* (also known as the Ishikawa or fishbone diagram) (Fig. 3). This tool is essentially a means of brainstorming in categories. The effect is at one end of a main “spine,” and categories of causes comprise the ribs. Typical categories include methods, machines, manpower, and materials, although other categories are often used. Observing the relationship between the entries can often assist in determining the root cause of the problem being displayed.

2. *Check Sheet* (Table 2). This is a simple method of displaying how often various events occur. Quick observation can help determine where to put effort to eliminate or minimize the largest part of a problem.

3. *Control Chart* (Fig. 4). The control chart displays counts or measurements in successive order over time (essentially a run chart) but includes the addition of statistically calculated control limits on either side of the process average. Using various run rules, the control chart allows you to differentiate between “common cause” variation (that is, the variation inherent in the system due to equipment, materials, processes, etc.) and “special cause” variation

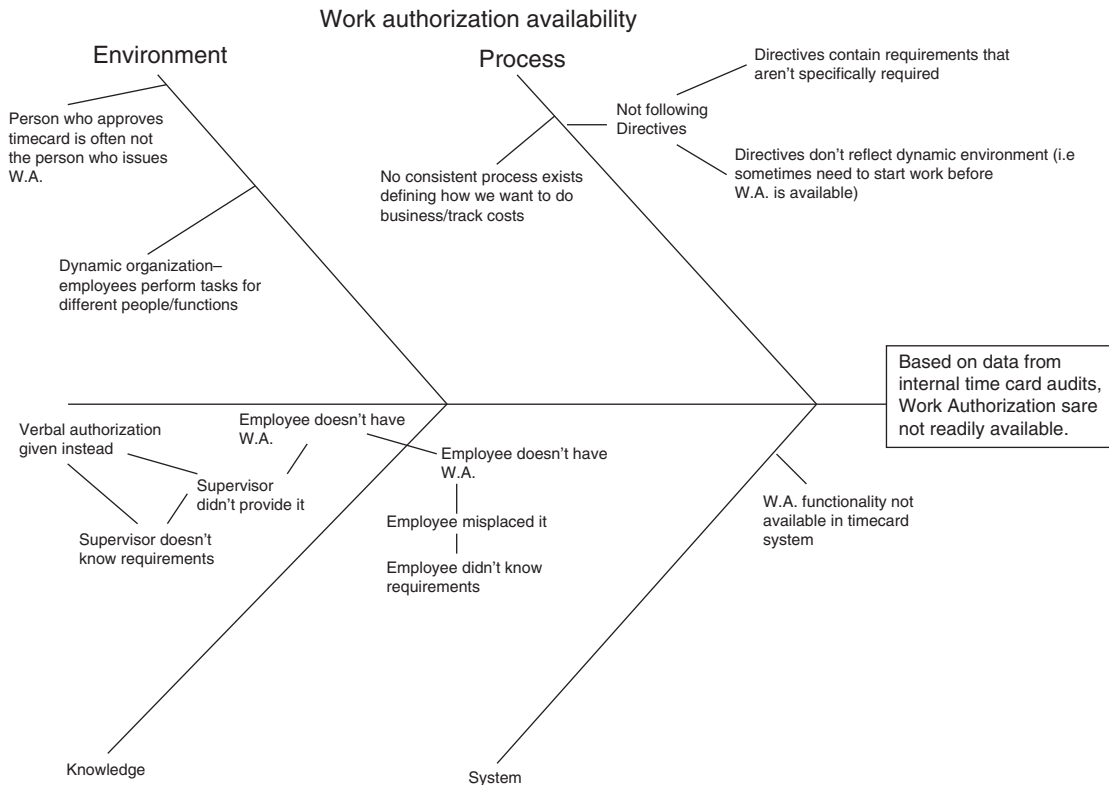


Figure 3 Cause-and-effect diagram.

Table 2 Check Sheet

Causes for Shipping Damage						
	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Dropped	//	///	////	//	/	12
Improper packaging	///	/		//	///	9
Water		///				4
Excessive heat				////		4
Vehicle accident	/					1

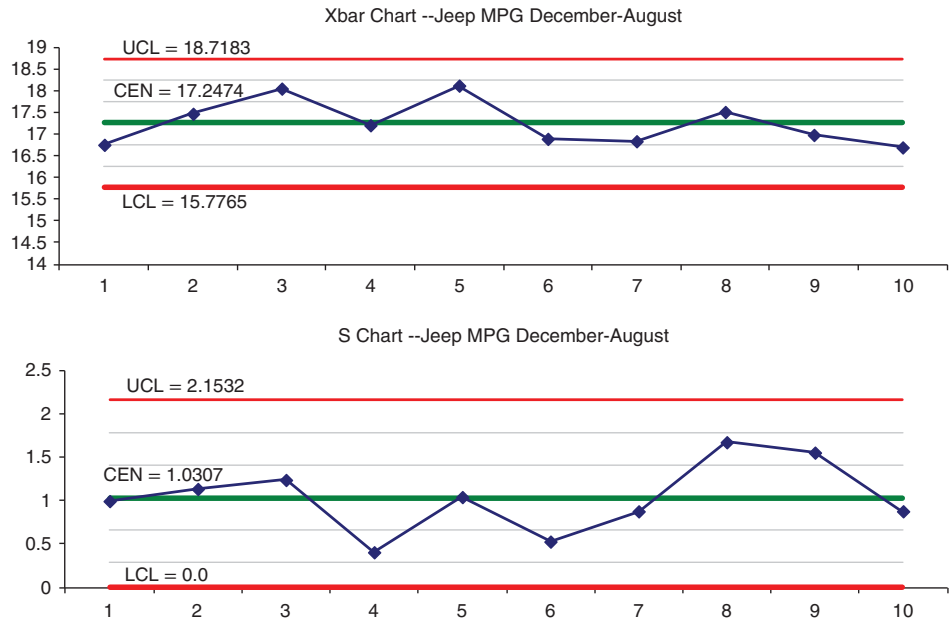


Figure 4 Control charts.

(the variation caused by errors, unpredicted events, etc.) Understand that being “in control” implies that the results are predictable. You can have an in-control process that is predictably producing defective product.

The control chart helps practitioner to understand when to take action and when not to. Based on the run rules, certain conditions require immediate response to fix and avoid future occurrences of the special cause problem. Common cause variation, or predictable process performance that produces unacceptable results, requires management intervention to eliminate or reduce the cause of the variation—design changes, machine improvements, improved processing methods, etc.

4. Histogram (Fig. 5). The histogram is a method to graphically display the frequency with which different events occur. It is essentially a formalized version of the check sheet. The histogram shows the cumulative variation in the process over a period of time but does not show trends. However, by understanding whether the variability is large or small in relation to the specification limits, whether variability is symmetrical around the process average, etc., will help to understand the nature of the process.

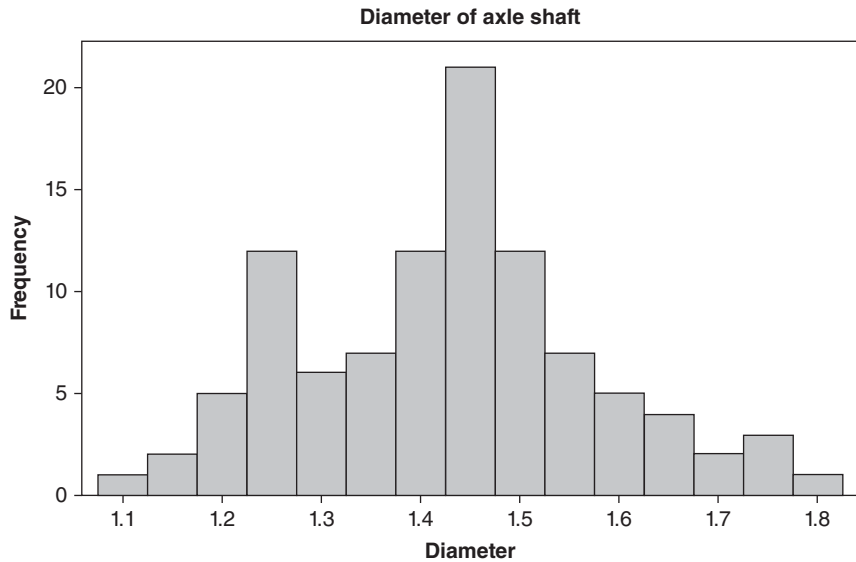


Figure 5 Histogram.

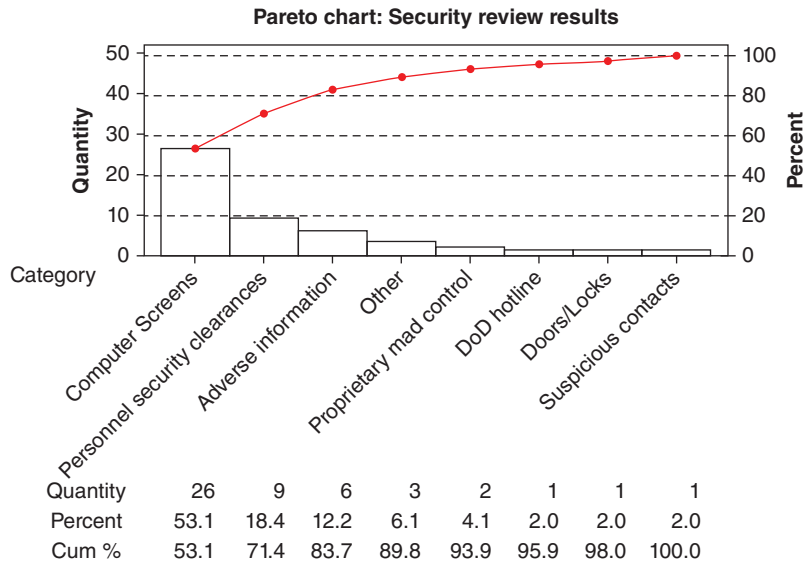


Figure 6 Pareto chart.

5. *Pareto Chart* (Fig. 6). The Pareto principle was developed by quality improvement expert Joseph Juran, named after Italian economist Vilfredo Pareto. Pareto had observed that 80% of Italian land was owned by 20% of the people. This spawned the 80/20 rule, which noted that in numerous examples, 80% of the effect is a result of 20% of the causes. Juran took that idea into the quality improvement arena, where he showed that in many cases, 80% of the problems were a result of 20% of the causes. In other words, by displaying your data in a Pareto

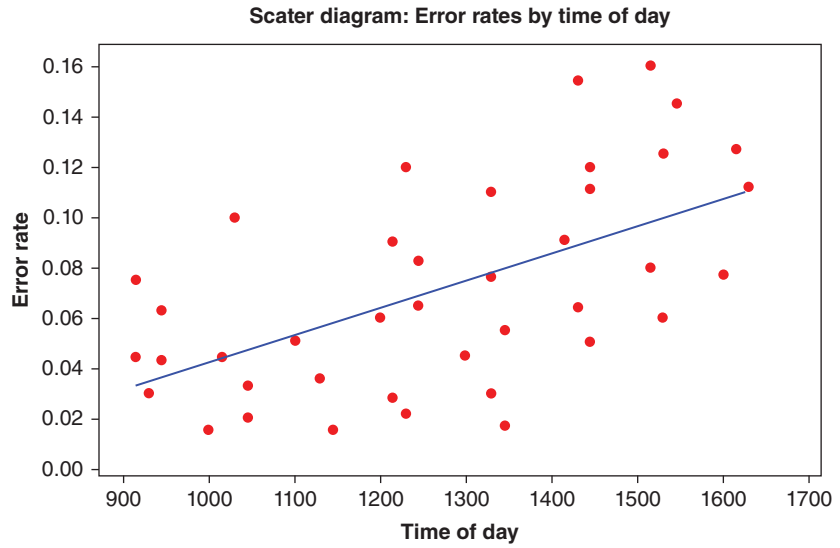


Figure 7 Scatter diagram.

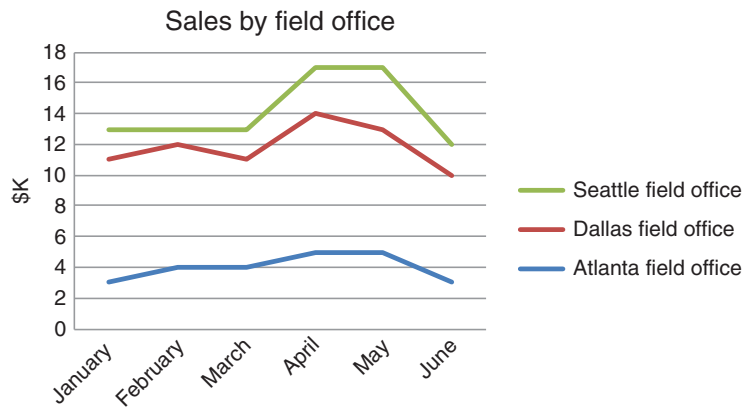
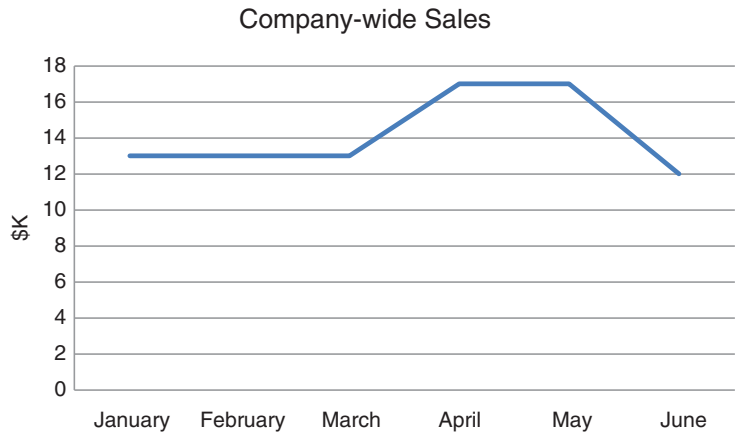


Figure 8 Stratification.

diagram, you can focus your improvement efforts on the causes that will achieve the greatest improvement.¹⁰

6. Scatter Diagram (Fig. 7). A scatter diagram offers an easy way to look for relationships between two variables. Note that it does not prove cause and effect, but it does show relationships.

The direction of the clustered points on the diagram can show whether there is a negative or positive relationship, and the tightness of the cluster indicates how strong a relationship exists. For example, if the points are tightly clustered in a straight line, a strong correlation exists between the variables—every time one variable changes, the other changes equally.

7. Stratification (Fig. 8). Stratification is a modification of a standard run chart. If a run chart combines several sources of data in the overall rate, it may not tell the entire story. By breaking down the causes into logical subgroups, a clearer indication of the situation is often shown.

6 SEVEN MANAGEMENT TOOLS

While the seven quality control tools are often usable throughout an organization and can be used by production floor personnel to address problems directly, a second group of improvement tools was developed in the late 1970s that proved valuable initially in Japan in drastically reducing cycle times. These seven management planning tools take complex tasks like planning production operations to the level where barriers can be broken and production floor personnel can participate in the process.

Again, the tools are not necessarily complex or difficult to understand.¹¹ Additional details and useful hints are available from numerous sources. The tools are as follows:

1. Affinity Diagram (Fig. 9). This tool takes a large group of data inputs and rearranges (affinitizes) it into related groups. Brainstorming ideas and recording them on Post-It® notes and then rearranging them into categories are a simple way to illustrate these relationships.

2. Interrelationship Digraph (Fig. 10). If you have a large amount of information and you need to understand the cause-and-effect relationships between the various parts, the interrelationship digraph can be useful in finding root causes. Information can come from other tools, such as affinity diagrams and cause-and-effect diagrams. For each entry, you ask if it causes or influences each of the other entries, going through all of the possibilities and drawing an arrow to indicate that cause or influence. When complete, an entry with numerous arrows either pointing to it or leading out of it indicates a possible target for further investigation.

3. Tree Diagram (Fig. 11). If you need to understand the component parts of a broad objective, a tree diagram can be a very useful tool. (The broad objectives can come from the major categories of an affinity diagram.) You state the broad objective on a Post-It® note and then determine what needs to happen to address that situation. Record the answer(s) on another Post-It® note and continue the process until the necessary level of detail is reached. A final check for logical flow is important. In addition, checking to make sure that all the tasks are necessary can help to streamline the operation by eliminating those that are not necessary for success.

4. Prioritization Matrix (Fig. 12). If you have identified key issues but must select the most advantageous activities to undertake, the prioritization matrix can overcome disagreement. Each activity is weighed against a series of criteria, which can be weighted to show relative importance. Numerous ways exist to construct the prioritization matrix, but one

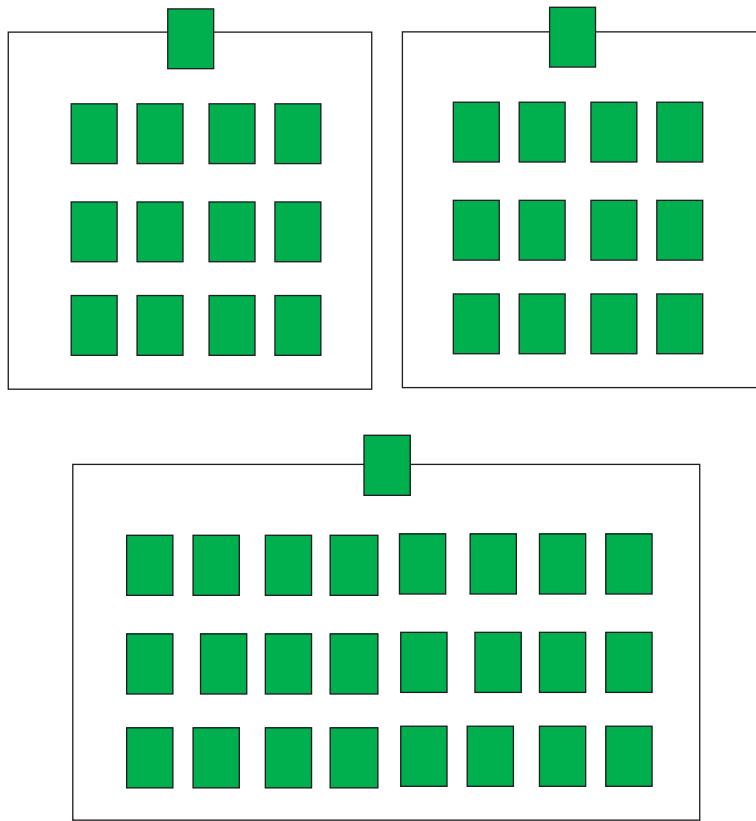


Figure 9 Affinity diagram.

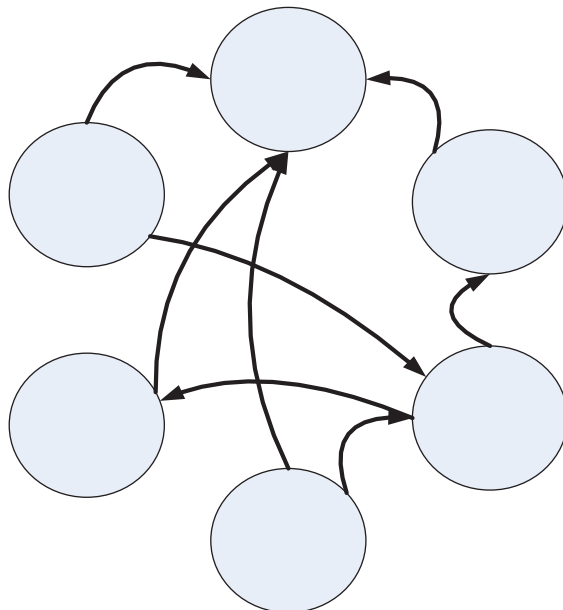


Figure 10 Interrelationship diagram.

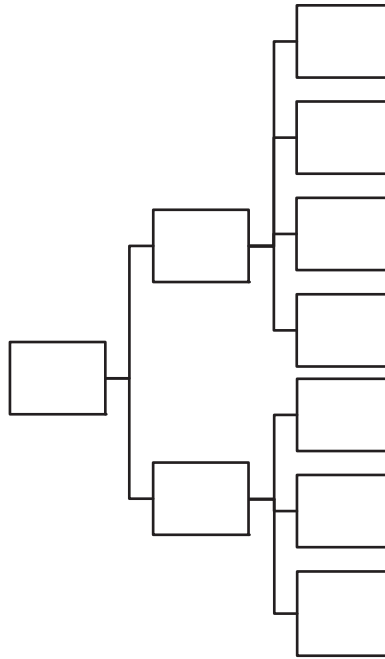


Figure 11 Tree diagram.

effective way is to number each relationship as 1, 3, or 9 (with 9 being the most desired outcome and 1 being the least desired). The entry in each box is multiplied by the weight applied to each criterion, and then each row is added. The row with the highest total is the one with the biggest impact on the problem and should be undertaken before those of lesser value.

5. Matrix Diagram (Fig. 13). The matrix diagram is used to show the relationship between pairs of items. Various chart layouts are possible, depending on how many items are being compared. A simple example compares items in the first column with other items in the top row of a square grid. The intersecting space is coded with symbols indicating a strong relationship, some relationship, and a weak or possible relationship. Any shapes can be used for the coding, but they should be easily differentiated.

6. Process Decision Program Chart (PDPC) (Fig. 14). The PDPC provides a variation on the tree diagram in which process steps are appraised for risk. A PDPC can be most useful in cases where a complex task with high stakes for failure is being developed or heavily modified. At each step, you assess what could go wrong, brainstorm possible solutions, and chart them on the PDPC.

Alternately, steps may be shown in an outline format, with substeps indented. Again, each step is assessed for what could go wrong and what alternative solutions exist.

7. Activity Network Diagram (Fig. 15). When planning a complex project, understanding the sequence of steps that will take the longest time to complete is essential. This sequence is known as the critical path, and completing these steps successfully drives the success of the overall project. Various disciplines have arisen from this tool, including critical chain project management, PERT charts, etc.¹¹

Criteria	Activity								Total	Rank
		Importance to community	Cost	Identifiable leadership	Time required	Minefield?	# of people impacted			
Weighting of Criteria		3	9	9	3	3	3			
1	Develop collaborative approach to solving water / waste water issues	9	9	9	3	1	9	225	1	
2	Develop spreadsheet of metrics & benchmark	3	9	9	3	9	1	183	2	
3	Public awareness Strengths Activities of various groups Entrepreneurial strengths of Tucson Water availability Etc.	9	3	9	3	9	9	171	3	
4	Develop efforts to retain graduates / attract those from elsewhere	9	3	9	3	9	9	171	3	
5	Participation in 2040 transportation plan	9	3	9	1	9	9	165	5	
6	Educate and promote benefits of collaboration	9	9	3	1	9	9	165	5	
7	Get part-time residents involved in process, \$	9	3	9	3	9	3	153	7	
8	Leadership training / seminars / boot camps	9	3	9	1	9	3	147	8	
9	Mediate dialog between public & private sectors	9	3	3	1	3	9	111	9	
10	Resource center for small business issues	3	3	1	3	9	9	81	10	
Meaning of 9 rating:		High	Low	High	Low	Not likely	High			
Meaning of 3 rating:		Medium	Medium	Medium	Medium		Medium			
Meaning of 1 rating:		Low	High	Low	High	Very likely	Low			

Figure 12 Prioritization matrix.

	A	B	C	D	E
1					△
2	△	○		◎	
3			○		
4		△		◎	○
5	◎		△		○

◎ = Strong relationship
 ○ = Some relationship
 △ = Weak / possible relationship

Figure 13 Matrix diagram.

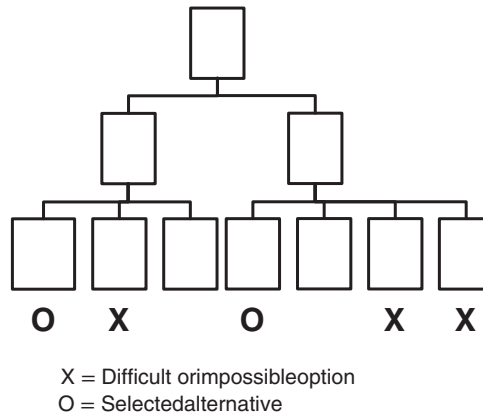


Figure 14 Process decision program chart.

7 LEADERSHIP AND LEAN

U.S. companies, particularly manufacturing, are fond of implementing “the next big thing” in improvement tools and techniques, realizing some quick improvements in rates or profits while seldom sustaining the effort to the extent that the changes become internalized. Employees often complain about the initiative of the day, and with each new improvement program, they become more cynical and dispirited. Lean manufacturing has been one of those initiatives since the 1980s with components such as just-in-time manufacturing, kanban, or Kaizen implemented, while the full program is not embraced by leadership and the tools that are used to provide immediate benefits are not sustained.

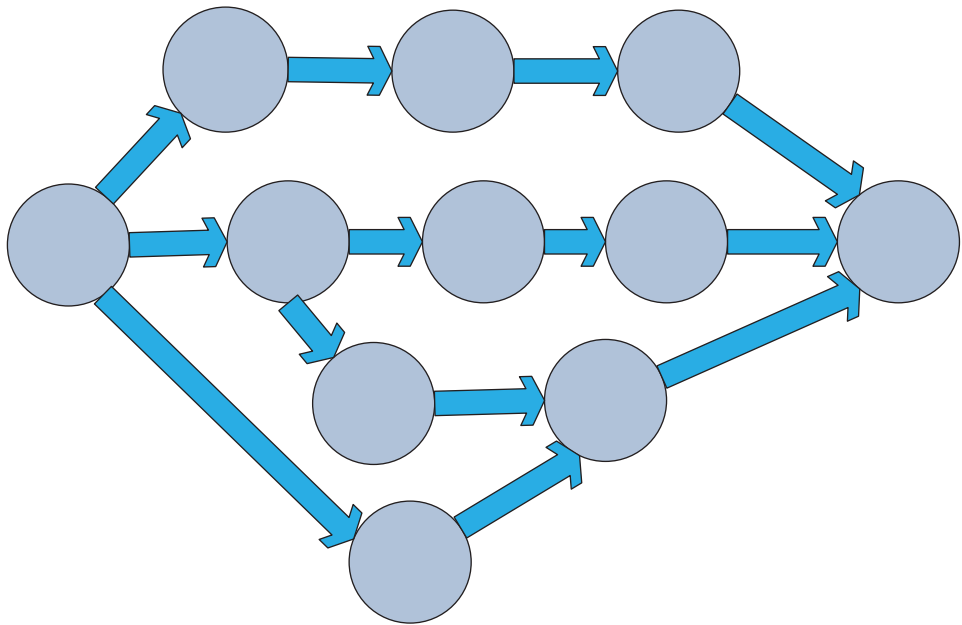


Figure 15 Activity network diagram.

There are many examples, however, of a single poka-yoke solution that continues to eliminate errors and rework even years after its implementation. So, in many cases, the use of the lean tools provides lasting improvements. It will be up to each management team to determine the scope of long-term benefits it is willing to attain and the extent of leadership changes it is willing to make.¹²

REFERENCES

1. N. J. Sayer and B. Williams, *Lean for Dummies*, Wiley, Hoboken, NJ, 2007.
2. T. Ohno, *Toyota Production Systems: Beyond Large Scale Production*, Productivity Press, Portland, OR, 1988.
3. J. P. Womack and D. T. Jones, *Lean Thinking: Banish Waste and Create Wealth in Your Corporation*, Simon and Shuster, New York, 1996.
4. T. Pyzdek, *The Six Sigma Handbook*, McGraw-Hill, New York, 2003.
5. D. W. Benbow and T. M. Kubiak, *The Certified Six Sigma Black Belt Handbook*, ASQ Quality Press, Milwaukee, WI, 2005.
6. B. Vernyi and T. Vinas, "Easing into E-Kanban," *IndustryWeek*, November 11, 2005.
7. T. Peters and R. H. Waterman, *In Search of Excellence*, HarperCollins Publishers, Inc., New York, 2004.
8. M. Brassard and D. Ritter, *The Memory Jogger II, A Pocket Guide of Tools for Continuous Improvement and Effective Planning*, Goal/QPC, Salem, NH, 1994.
9. J. M. Juran and J. DeFeo, *Juran's Quality Handbook: The Complete Guide to Performance Excellence*, 6th ed., McGraw-Hill, New York, 2010.
10. M. Brassard, *The Memory Jogger Plus+*, Goal/QPC, Methuen, MA, 1996.
11. L. P. Leach, *Critical Chain Project Management*, Artech House, Norwood, MA, 2005.
12. J. K. Liker and G. L. Convis, *The Toyota Way to Lean Leadership*, McGraw Hill, New York, 2012.

CHAPTER 22

TOTAL QUALITY MANAGEMENT FOR MECHANICAL ENGINEERS

Alan Kemerling, PhD
Ethicon, Inc.

1 TWO PATHS: DMAIIC AND DMADV	636	3.3 Analyze Phase in DMADV	660
2 DMAIIC	637	3.4 Design Phase in DMADV	661
2.1 Define Phase in DMAIIC	637	3.5 Verification and Validation Phase in DMADV	663
2.2 Tools for Define Phase in DMAIIC	637	4 SUMMARY	664
2.3 Measure Phase in DMAIIC	639	REFERENCES	664
2.4 Analyze Phase in DMAIIC	643	REFERENCES NOT CITED	664
2.5 Improve/Innovate Phase in DMAIIC	649	RECOMMENDED FURTHER READING	665
2.6 Control Phase in DMAIIC	653		
3 DMADV	656		
3.1 Define Phase in DMADV	656		
3.2 Measure Phase in DMADV	657		

Why should a mechanical engineer be interested in total quality management (TQM)? One reason is that the processes developed to support TQM offers the engineer a proven way to do projects with the highest possibility of success. Mechanical engineers often lead project teams in the development of new products or processes. They also lead teams in the improvements of products and processes. Engineering class work may not prepare you for these roles. The processes in TQM provide paths that not only work but also can be understood by the wider audience of team members and upper management. See the key takeaways below.

TQM offers proven processes and tools for success in:

- Process improvement
- Product improvement
- Product or process development

The second reason to use TQM principles is that they work. Time and again, in organization after organization, the practices of TQM provide their businesses with results that help their customers and the bottom line. The two go together. A measure of the success of TQM can be seen in the winners of the Malcolm Baldrige Performance Excellence Award administered by the National Institute of Standards and Technology. With categories in education, health care, manufacturing, nonprofit/government, service, and small business, those who work with the Baldrige process apply TQM principles and tools. When you go to the website (www.nist.gov)

and look at the Baldrige winners, you will find hospitals that are reducing adverse event rates, increasing patient satisfaction, and increasing staff satisfaction while they reduce costs. You will find education, nonprofits, and government agencies improving their reach to those who need their services while holding costs down. You will find manufacturers reducing costs and improving deliveries while increasing profitability.

Recent results from the 2011 Malcolm Baldrige Award winners demonstrated the breadth of the application of TQM.¹ Winners included a small publishing house in St Louis and three health care facilities in Michigan, Indiana, and Alaska. Health care has been a rising applicant of TQM for some time because the competing forces in health care and the importance of its mission have been mandating improvements to control costs and improve services. A look at the stories from these organizations shows improvements in service, profitability, and personnel retention. At the same time, one can find articles noting cases where TQM ventures did not have the promised payoff. When that happens, there is always a clear reason for the failure. As noted in a recent *Wall Street Journal* article, the reason for failure is not the principles or tools.² The reason for failure is the lack of support by management. The mechanical engineer leading an improvement or development project must be able to express his or her needs to management to make sure the project is successful.

In 2010, there were seven Malcolm Baldrige Award winners. One of the manufacturing winners, MEDRAD, demonstrated growth of its improvements per employee from \$23,000 in 2005 up to \$45,000 in 2009. This was one of several metrics showing improvement of this company. Freese and Nichols, an engineering consulting firm, demonstrated revenue growth between 12 and 16% from 2005 to 2009. It was also awarded one of the top 25 companies to work for in 2009.

For the mechanics of TQM, this chapter will make the case for using the two different flows or processes of TQM, depending on what problem you are facing as an engineer. Along the steps of the process, you will be introduced to some of the tools that may be used. The large number of tools and the complexity of some of them preclude extensive detail on each, but the basics will be presented and other resources will be given at the back of the chapter. The vast resources of the Internet will give you more detailed assistance as you find need to learn more about a particular tool.

It should be noted that TQM is often known by other names, including Six Sigma, Process Excellence, Design Excellence, and Lean Six Sigma. The processes presented here serve as the base for all the variants.

1 TWO PATHS: DMAIIC AND DMADV

There are two basic paths in TQM, one is called DMAIIC (pronounced duh-may-ick) and one is called DMADV (pronounced duh-mad-vee). These acronyms stand for the DMAIIC process steps of define, measure, analyze, improve/innovate, and control and the DMADV process steps of define, measure, analyze, design, verify, and validate. DMAIIC is sometimes written as DMAIC, where the single I stands simply for “improve.” Including the additional “innovate” term is preferred by this author as a way to indicate that thinking of improvements may not be enough for your business. Sometimes you need to think in break-through terms instead of just incremental improvement. The term DMAIIC will be used in this chapter.

These processes start in a similar manner, but there are distinct differences. To decide which process to employ, it is important to understand what you need to accomplish. Stephen Covey³ always encouraged people to “Begin with the end in mind.” This is good advice here also. A DMAIIC process will *typically* be employed if there is an existing process that must be improved. Here the term *existing* is used loosely. You may have an existing ad hoc process that is generally followed but not written out anywhere. Such a process is generally a candidate for DMAIIC. A DMADV process will typically be employed in product or process development

Table 1 Two Paths

DMAIIC	DMADV
<ul style="list-style-type: none"> • Define • Measure • Analyze • Improve/innovate • Control 	<ul style="list-style-type: none"> • Define • Measure • Analyze • Design • Verify • Validate
Use this path for improvement of products or processes, even if the predecessor was ad hoc.	Use this path to develop new products or processes. This can also be used if the redesign is extensive.

where new development is required because there is nothing existing or because the existing situation requires extensive or breakthrough improvement. See Table 1 for a summary.

In the rest of this chapter, the DMAIIC and DMADV processes will be detailed and some of the tools that may be useful will be listed and explained. The rest of this chapter will assume that the mechanical engineer is leading a team for the project, although these processes may also be used by an individual very successfully.

2 DMAIIC

As you move through the phases of the project, it is very helpful to think in terms of questions that need to be answered. The questions define what is needed in the next steps and provide guidance on the tools to employ. The following sections will include examples of key questions to ask.

Before you start improving a process, the first question that should be asked is, “Should the process be retained at all?” It is possible that revised technology and customer expectations allow you to do away with the process completely. An example is the idea of checkout at some stores. Most stores have improved the existing process for the checkout person through the use of scanners and automated scales, but stores are now replacing the process with self-checkout. This is an example of rethinking the process rather than improving it.

2.1 Define Phase in DMAIIC

The first step is fairly simple but exceedingly important. The question that needs to be answered here is, “What is it that I or my team needs to do for the business or our customers?” In the *define phase* the team will detail what needs to be accomplished, and it is especially important that there be an agreement between the team and management on the required changes, timeline, resources, and budget of the project. Some projects can span a significant amount of calendar time, and memories of the original agreement can differ when the team is ready to wrap the project up, but management thinks there was more to the project. The define phase determines what the end looks like. It can also be described as defining what “done” will look like (how do you know you have finished).

2.2 Tools for Define Phase in DMAIIC

Charter. The key tool to use in the define phase of the project is a charter. The charter spells out what will be done, the timeline, and budget. There is another reason to employ the charter.

There are times when an improvement effort in one process or project will expose other things that need to be “improved.” When this happens, there may be pressure to fix these issues also, to the detriment of the original effort. It is important for the success of the team to avoid trying to “boil the ocean.” This is accomplished by putting some boundaries around the original project. New tasks can certainly be added if the sponsors agree it is the right thing to do, but allowing the scope of the project to creep larger and larger puts the original effort at risk. See the discussion of the in-scope/out-of-scope tool later in this chapter. It is useful to manage issues as they come up.

The format that the charter takes is not really important. The overall goal will be spelled out in SMART form. SMART stands for specific (well defined), measurable (not vague), achievable (realistic, within the reach of the team and its resources), relevant (in line with business needs), and timely (achievable in a timeline that is useful to the business and its customers). An example charter that you can use is at www.asourceofquality.com/resources.

SIPOC. When you are working on a process, SIPOC (which stands for supplier, input, process, output, and customer) allows you to develop a level of understanding for your team and to explain the project’s aim. It is also a useful tool for communication beyond the project team and the rest of the organization. It is used to make sure key aspects of a process are not overlooked in the investigation. If you are doing an improvement as part of a multifunctional team, this is a great tool to ground everyone in the major parts of the process. It is also a great feed into the next part of the process, stakeholder analysis. See www.asourceofquality.com/resources for an example of a SIPOC, but do not worry too much about the form. See Fig. 1 for an example.

Stakeholder Analysis. It may be easy to misunderstand the extent of communication necessary for a project. Poor communication can make or break the success of it. Stakeholder analysis helps determine who you need to communicate with and what the outcome needs to be. In a simple stakeholder analysis, list all the potential stakeholders of a project, both those who are involved with the project as well as suppliers and customers. Determine their anticipated support and your strategy to confirm or change their support level prior to the project. An example stakeholder analysis is available at www.asourceofquality.com/resources.

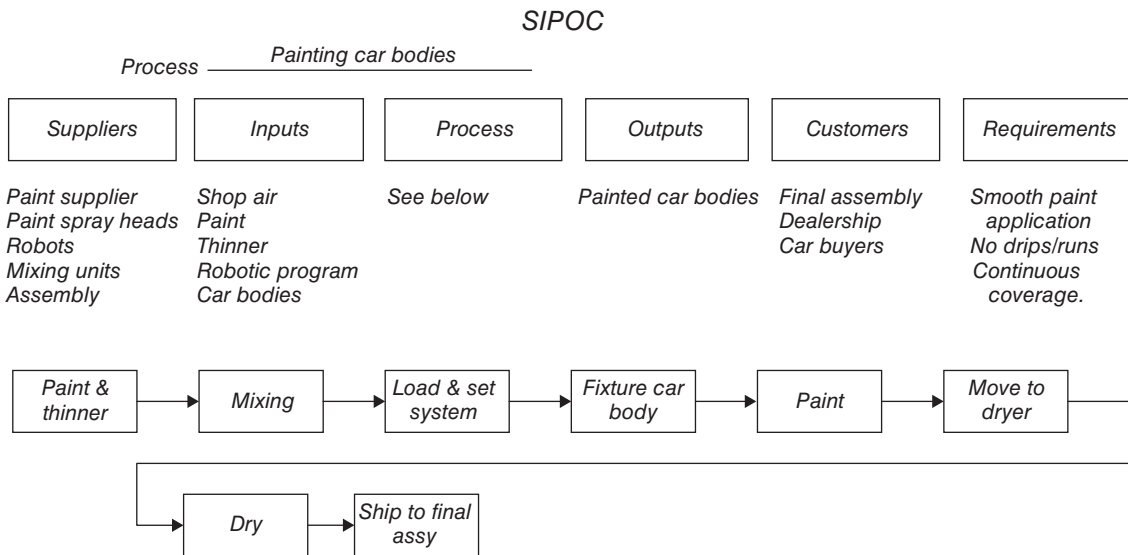


Figure 1 SIPOC.

Quality Function Deployment (QFD). Quality function deployment is a major tool that is used in both the DMAIIC and DMADV processes. It will be explained in the DMADV section of this chapter.

2.3 Measure Phase in DMAIIC

The key questions in the measure phase are, “Where is our current performance?” and “Where does it need to be?” The measure phase of the project is where the team will obtain and begin to analyze the data necessary to drive the project to completion and then close the loop on testing the improvement. If you want to think in the form of a mathematical approach, think of the inner part of the DMAIIC process as solving for $y = f(\mathbf{x})$, where y is some output of interest and \mathbf{x} is a set of input variables. The first thing we need to do in the measure phase is find y_{current} and y_{desired} (if the desired end point was not supplied in the define phase).

Tools for Measure Phase in DMAIIC

Data Collection Plan. If you are fortunate, the data are already available for your project. If not, you must start with a data collection plan. The data collection plan is very simple and the format does not have to be very specific. A table developed in Microsoft Word would be sufficient to share with the team. Some of the columns would be Data Description, Type of Data, Measurement Method, Criteria, Measurement Point, Sample Rate, and Data Collection Form(s). See www.asourceofquality.com/resources for a simple example of a data collection plan.

Data Collection Forms. If the data do not exist and the team has to obtain data, data collection forms may be developed to assist in the effort. Make sure each data collection form spells out the following:

- What data are to be collected (what is to be measured/observed)?
- How will it be collected (tools, gages, fixtures, etc.)?
- How often will it be collected (what point in the process and condition)?
- Where will it be recorded?

Observe the data collection effort to see if it is going as you anticipated so that the data you ultimately use are accurate.

Measurement Systems Analysis (MSA). This used to be called gage repeatability and reproducibility (GR&R), but it applies to more than just gages. Since everything has variation at some level, it is no surprise that measurement systems themselves have variation also. Variation in a measurement system comes from three sources. The following talks about *gages*, but it is important to note that you still have a measurement system if you have an inspector or classifier.

- *Bias in the Gage*. If you employ a gage, it is important to know that all gages have bias. When a gage has a bias, it tends to indicate a reading that is above or below the true value. For gages, this is a function of the gage’s calibration. For operators and inspectors, it is a matter of training and good examples or inspection aids.
- *Repeatability*. When a gage is not repeatable, it means that repeated measurements by the same operator, when the part is removed and replaced in the fixture or gage, show a certain variation. Repeatability is influenced by gage design. Electrical noise, excess play in mechanical linkages, or a loose fit in fixture retaining features can influence repeatability. For operators and inspectors, it is a matter of training and good examples or inspection aids.
- *Reproducibility*. This pertains to the ability of a second operator to achieve the same result as a previous operator working with the same equipment and under the same

conditions. Examples of factors that influence reproducibility include holding fixtures that are sensitive to operator technique and measurement instructions that give the operator significant discretion in how the part will be mounted and measured. For operators and inspectors, it is a matter of training and good examples or inspection aids.

Some may be surprised at this level of detail for process measurement systems, but the driver is economic. If a process has lower than desirable yield, the issue may be with the process, measurement system or both. The measurement system may be rejecting good parts and allowing nonconforming units to be sent to customers. There are three good reasons for attention to detail in measurement. First, it is often more economical to fix a measurement system than change a process. Second, if the measurements being taken have a large amount of uncertainty; it is likely that you are rejecting good parts, delivering parts that are not in conformance, or both. Third, if there is a large uncertainty about measurements, it may be difficult to know if process improvements were successful.

Measurement systems analysis should be performed properly so the source of variation is identified. It is desirable that the parts used in measurement systems analysis exhibit variation covering the expected tolerance range, although this may be difficult for some processes. Most analysis involves approximately 10 parts and two to three operators. First, the gage is calibrated or the calibration record is checked. Second, each operator will measure and record the features of interest on each part two or three times. Parts will be run in random order to remove any time trending with the operator or measurement system. All measurements will be recorded with identification of the operator, part, and order of measurement. The statistical analyses will then break down the sources of variation and identify how much variation is coming from the parts, the repeatability of the gage, and the reproducibility of the gage. See <http://www.itl.nist.gov/div898/handbook/mpc/section4/mpc4.htm> for more information on gage R&R studies.

Many companies place guidelines on the amount of measurement error they will tolerate in the system. Generally, less than 20% of the feature tolerance is an acceptable range. If the error is less than 30%, it may be tolerable depending on the criticality of the feature. If the measurement error is greater than 30% of the tolerance range, the measurement technique should be improved.⁴

Another aspect of measurement systems analysis is the comparison between gages. Often companies will rely on suppliers' measurements. If there is an issue, it is good to be able to assess parts at your facility and know your measurements will be similar to your supplier's.

Also you need to ensure that the gage is appropriate for the type of measurement taken. If you need three significant digits to the right of the decimal point in measurement, make sure the resolution of the gage is adequate to supply this type of measurement.

Process Capability Analysis. Process capability studies allow engineers and operators to put an estimate on the long-run or short-run performance of the process. Knowledge of process capability may aid in setting specifications or support the prediction of scrap, rework, and throughput. If design engineers understand process capability and use that information to set specifications, there can be less conflict between design and manufacturing. Process capability takes on several forms, but the primary form is noted in the literature as C_{pk} .

For many companies, engineering design has been slow to understand the need to work with manufacturing to create a design package that meets customer needs and can be manufactured economically. For their part, manufacturing has not always been proactive in developing consistent processes with minimum variation and communicating process requirements and capabilities to design engineers. There is plenty of blame to go around, so how do we change? A key way to change is to look at facts and data. If you are supporting manufacturing, characterize your processes and communicate process capabilities to designers and external customers. If the design requires certain tolerances but the process cannot maintain that performance, the

only thing that may be done is to change the process! Otherwise, the people supporting the process will always be fighting poor yield and the losses must be reflected in part prices. The following are steps to follow:

1. Prioritize your processes according to highest loss (scrap, rework, cost, etc.) and start working on the highest ones (the vital few).
2. If the process does not have SPC (see discussion of SPC under the control part of DMAIIC), apply it!
3. Get the process under statistical control, that is, predictable.
4. From the SPC chart, obtain estimates of the process average and standard deviation.
5. Assess the process C_{pk} (see below).
6. Based on the resulting C_{pk} , determine whether to
 - 6.1 change the product specification or
 - 6.2 improve the process using the DMAIIC process.
7. Move to the next process in the list.

In step 1, develop a comprehensive strategy. Many organizations go after the processes with the most scrap or the most overall cost. In steps 2 and 3, stabilize the process by removing sources of special cause variation. In steps 4 and 5, use existing data from a stable SPC process to assess capability. In step 6, determine the best approach for your business. Assuming the process performance is not acceptable, determine your next strategy as follows. If the stable process capability is low but the product specification may be easily changed, the cost of an engineering change is nearly always less than a process improvement effort. If the process performance is unacceptable, process improvement may be warranted. The last step calls for the team to move on to the next process on the list, driving for continuous improvement.

The capability index C_{pk} indicates how much room there is between the product specification (tolerance) limits and the expected output of the process. The C_{pk} calculations and performance values are given in Fig. 2, where C_{pk} shows how many multiples of three standard

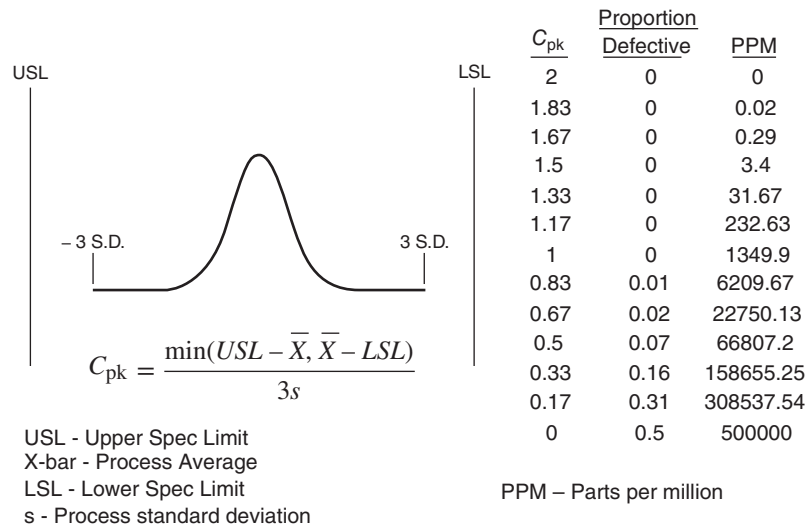


Figure 2 Capability.

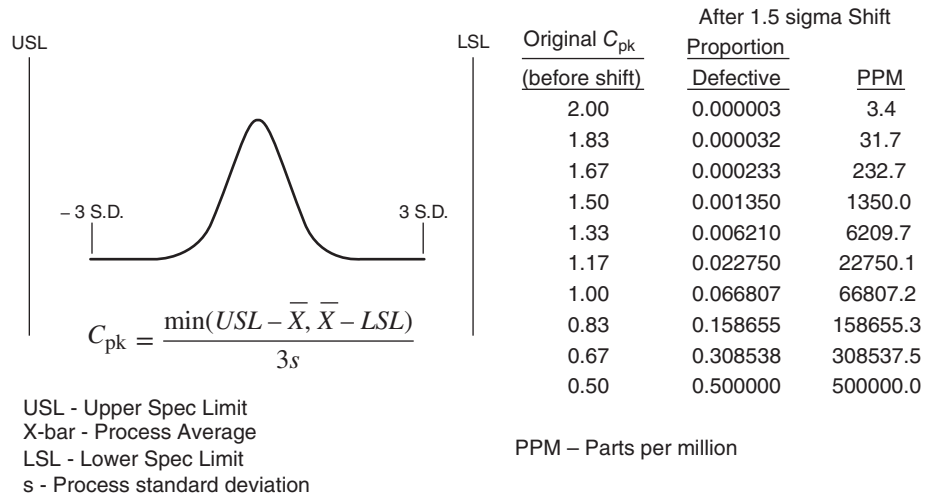


Figure 3 Shift.

deviations fit between the process output average and the closest specification limit. A C_{pk} of 1.0 indicates there are only three standard deviations between the process average and a specification limit. A C_{pk} of 1.33 indicates that there are a minimum of four standard deviations to the closest specification limit. Your company may have established a target value for C_{pk} . Some companies use 1.33 or 1.5 as the target value. Higher values of C_{pk} allow greater margin if the process slips out of statistical control. You can see what happens in Fig. 3 if the process C_{pk} slips from 1.5 to 1.0.

Steps 6 and 7 are very important. If the process capability is not acceptable, the design of the device or the process must be changed. If one or both of these are not done, your business must live with the resulting low process performance or poorly performing design as long as the product is made. The decision of which to address—product design, process, or both—is an economic one. When you have completed all the steps for one process, move to the next one.

Another aspect of process assessment is the measurement system. If a significant portion of process variation comes from measurement, it may be easier to improve the measurement system than the process. Gage repeatability and reproducibility assessments are also a core current part of the process.

There is a methodology often mentioned in the literature called the Six Sigma process. This process would guide the team to set a process and product specification so that there are six standard deviations between the process average and the closest specification. A Six Sigma process would have a C_{pk} of 2.0 or greater (check the calculations to ensure this relationship is understood). Six Sigma strives for the extra margin because a process will often encounter a shift. It is possible that process shifts of up to 1.5σ may occur. This is the amount of shift that may occur before the change is detected and corrected. If the process is six standard deviations from the nearest limit and a process shift of 1.5σ occurs in that direction, the engineer will still have a process operating at $1.5C_{pk}$. Figure 3 shows the effect of a 1.5σ shift under various C_{pk} values. Machine wear, setting up the process incorrectly, new operators, changes in material batch, and other issues can cause process shifts.

The concept of process shift is incorporated into a variation of process capability called process performance. Process performance, labeled P_{pk} , assesses how well the process output conforms to the specification over a longer time span. While a process capability might be assessed one time using a single sample, process performance might be assessed over several

months. Such a time would expose the output from several operators, shifts, and batches of raw materials. There are statistical ways to obtain P_{pk} from control chart measures. Consult process control chart references in the back of this chapter for more information. As with C_{pk} , your company may establish target values for P_{pk} . Often, lower target values for P_{pk} are tolerated because process shifts are expected and should be detected and corrected by process controls. For example, a company may have a minimum target C_{pk} of 1.5 and a minimum P_{pk} target of 1.0.

In-Scope/Out-of-Scope. This tool can take several forms from a list to a graphic on the wall. The best form might be a simple spreadsheet that is reviewed periodically with the sponsors and decisions can be recorded along with the date they are made. See www.asourceofquality.com/resources for an example.

2.4 Analyze Phase in DMAIIC

After you have determined $y_{current}$ and $y_{desired}$ it is time to work on the set of x values in the process. The analyze phase is where you begin to evaluate the possible process values that have influence on the product or process. The important question for this phase is, “What do I need to do to drive $y_{current} \rightarrow y_{desired}$ in the most efficient manner possible?” This will involve some experimentation.

Tools for Analyze Phase in DMAIIC

Cause-and-Effect Diagram. Also called the Ishikawa diagram after Dr. Ishikawa, who introduced its usage, or a fishbone diagram from its distinctive shape (see Fig. 4), this chart helps a team identify the potential sources of a problem from what are often common process sources. These common sources are the material, machines, personnel (operators), measurement,

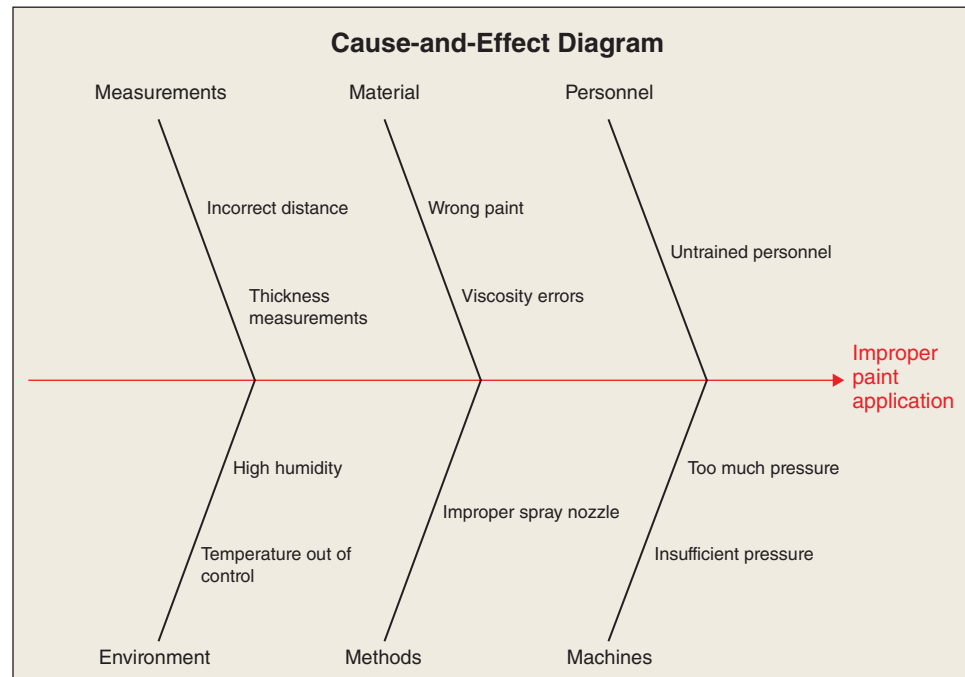


Figure 4 Fishbone diagram.

method (type of process), and environment around a process. The problem is noted on the main bar of the chart. The six possible sources are shown as diagonals from the main bar. The team then brainstorms specific sources to link to each of the six bars. A team may discuss and vote on the most likely source(s) of the problem for further analysis. This is a way for the team to whittle the possible sources of issues down to a manageable few for investigation.

The main purpose of the cause-and-effect diagram is to force the team to focus on all the possible aspects of a problem and then select the most likely source(s). By looking at each leg of the chart, material, machines, and so on, the team is asked to generate potential sources of the process issue from each aspect. This helps prevent the team from jumping to one solution and it can help keep one forceful person from dominating the discussion. It at least opens the mind to consider other possible sources. The fishbone diagram is a good lead into designs of experiments, which are discussed later in this chapter.

Graphical Data Analysis. Although they are simple, never underestimate the power of graphs to hint at what might be going on. Use graphs that are available in Microsoft Excel or more specialized programs such as Minitab, Statgraphics, SAS, or JMP. Take a look at the *Engineering Statistics Handbook* (<http://www.itl.nist.gov/div898/handbook/eda/section3/eda33.htm>) for examples and explanations on a number of graphical methods to explore data. For example, Fig. 5 shows what information may be obtained by observing simple plots. One plot shows a linear response, the other shows that one would expect some kind of exponential or quadratic response to this variable.

Histograms are often useful for variables data. The shape of the histogram will hint at the underlying pattern of data. If the problem may be time based, plot data by time or production lots. If there are multiple lots of components or materials, explore any resultant shifts by using plots that are grouped according to an underlying material or component lot. For example, Fig. 6 shows a difference between two suppliers.

Cosmetic defects may be charted by keeping a small diagram of the product and having operators put a mark where the defects are occurring. This might suggest where in the process the defect is being created.

Pareto Principle and Diagram. Dr. Joseph Juran brought to the attention of quality practitioners the fact that an ordered plot of counts of attribute data such as defect types very often showed a consistent pattern. Specifically, most process problems came from a relatively small set of sources (and hence generate common defect types). He suggested modeling attribute data in an ordered bar chart (largest count to smallest) to demonstrate this phenomenon. He named it the Pareto chart, after Vilfredo Pareto, a nineteenth century economist who noted such a pattern in Italian land ownership. Teams can use a Pareto chart of defects to focus on the problems with the most process impact. It is usually shown as a bar chart with a cumulative line graph overlaid on it. It is easily drawn using Microsoft Excel or other software programs with graphing capability. See Fig. 7 for an example. The Pareto principle is often expressed as the 80–20 rule. Generalized this would say that 80% of the problems you deal with will come from a small (~20%) of the different products you stock. For a particular line, a small number of defect types (~20%) will generate 80% of the defective product.

Design of Experiments and Hypothesis Testing. A key responsibility of a mechanical engineer is to obtain the required performance from a device, component, or process. This must also be done in the most efficient way possible for the company. This usually requires simulation, trade studies, or some form of experimentation with the possible input variables of one or more processes. Engineers are typically taught methods that include assumptions or approximations for the underlying equations. These may not be accurate enough to guide the engineer to the most efficient result.

Design of experiments (DOE) is the tool of choice for trade studies and design or process experimentation. A properly designed experiment will yield the most information possible from

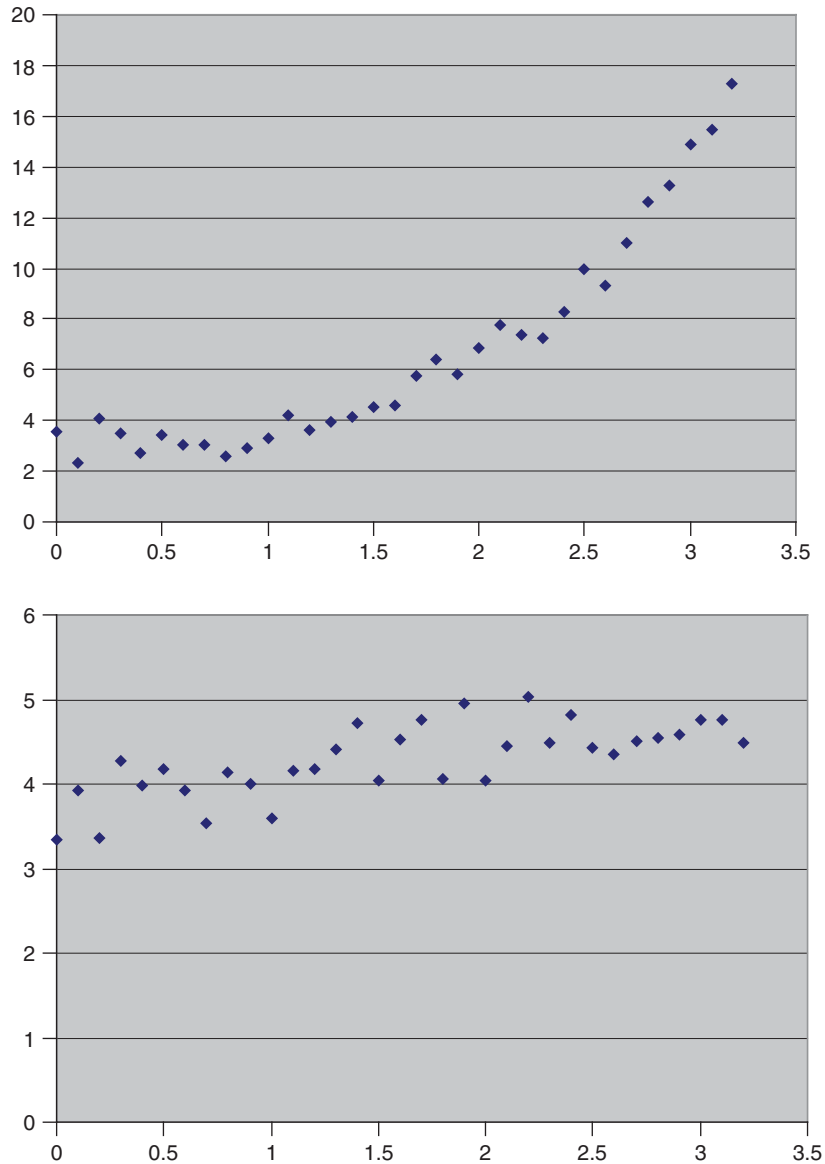


Figure 5 XY plots.

a given number of trials, fulfilling the engineer's fiduciary responsibility to the company. More importantly, properly designed experiments also avoid misleading results.

The chief competitor to good DOE work is the one-factor-at-a-time (OFAAT) approach where the engineer changes just one factor. This is repeated as the engineer works one at a time through all factors of interest while monitoring the response(s). OFAAT has some appeal because of its simplicity. Unfortunately, OFAAT yields only linear, first-order responses. The engineer often knows there are interactions between the factors or a factor's effect may be

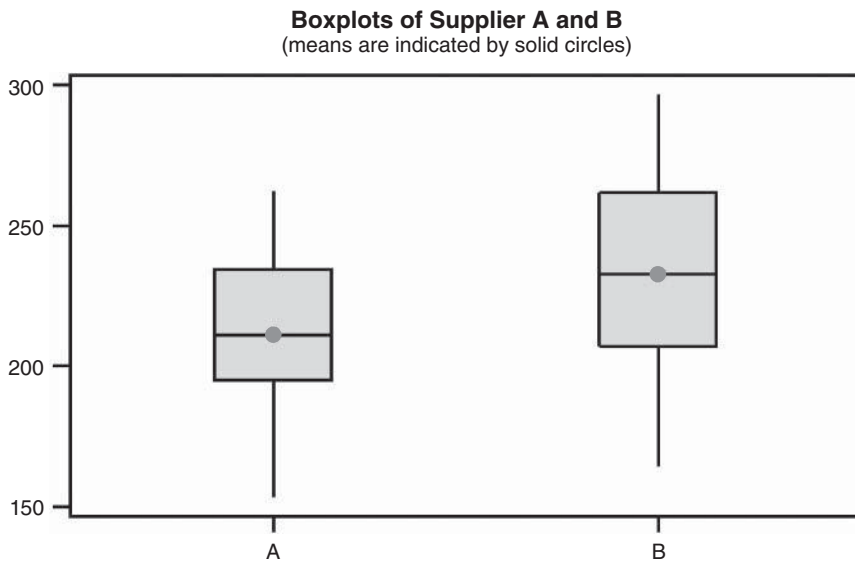


Figure 6 Box plots.

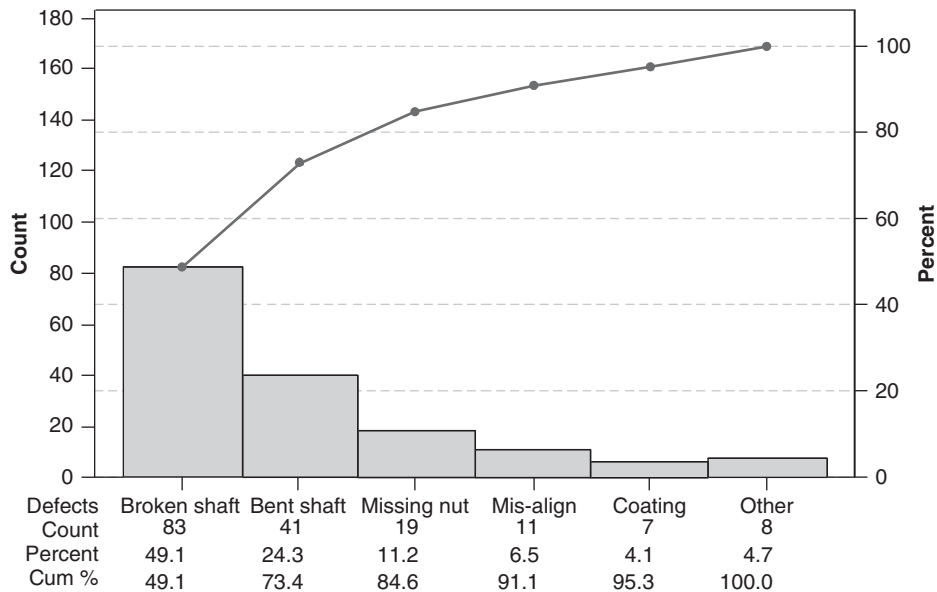


Figure 7 Pareto.

nonlinear (exponential or quadratic). OFAAT will not disclose this and could be very misleading. This approach is also described as a “Whack-A-Mole” approach after the carnival game where you try to hit little plastic moles as they pop out of holes in the game board. It is hard to keep up!

In Fig. 8, a system space is shown consisting of three factors, each at two levels. Experimenting with OFAAT will only explore the circled corners, yielding no information about the

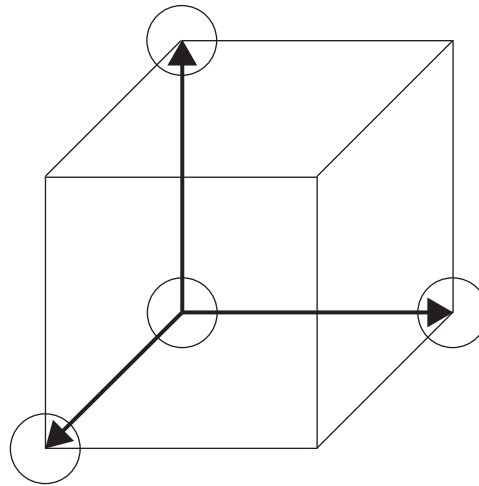


Figure 8 OFAAT.

remainder of the space. If there is any interaction between the factors, it would have shown in the corners that are not explored in OFAAT. We also do not know if there is any center curvature or if the response between corners is linear. To discover that would require some experimentation in the interior.

Another competitor to OFAAT is random experimentation. This takes place when the engineer changes more than one factor at a time, perhaps making multiple runs while trying different combinations. With random experimentation, a change in results may be stumbled upon, but it will be inefficient and the engineer will not know exactly why the improvement was achieved. The engineer may make a costly design or process change that is not necessary. Figure 9 shows a path of random experimentation. Like a random walk, this approach lacks an orderly approach to assessing the process environment.

As compared to OFAAT or random experimentation, well-planned DOEs systematically change factors according to a plan, measuring response(s) under known conditions. The

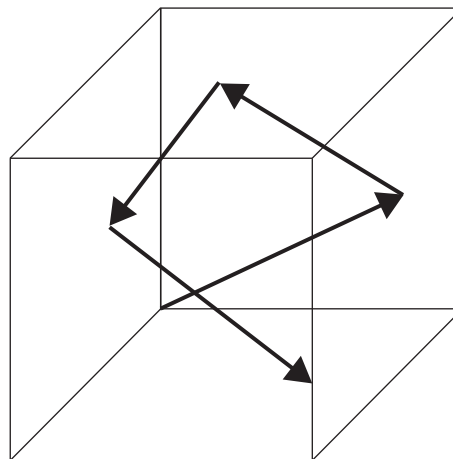


Figure 9 Random.

experiment often starts with a multifunctional team agreeing on the likely important factors for the experiment. The team may use a fishbone diagram coupled with voting to determine the priority of process factors. After determining what factors to use, the team must also decide how many levels for each factor. More factors and levels drive more experimental runs, so the factors and levels might need to be minimized. Initial experiments often keep the factors at only two levels. This helps reduce the number of experimental runs and makes the analysis somewhat easier. For example, an experiment with three factors at two levels and one factor at three levels will require $2^3 \times 3^1 = 24$ runs for just one replicate and one replicate is not usually sufficient.

There are many types of experimental designs, but they all fall into two major classifications:

- *Full Factorial*. An experiment where all possible combinations of factor levels are run at least once. If there are n factors, each at two levels, this will require 2^n runs for each replication. This type of experiment will yield all possible information but may be more costly than the engineer or company can afford. See Fig. 10 to see how all the corners of a three-factor space would be covered by a full factorial.

- *Fractional Factorial*. An experiment where a specific subset of the possible factor-level settings is run. A fractional factorial experiment only provides a subset of the information available from a fully factorial experiment. Even so, these designs are very useful if the subset available is planned carefully. Usually, a design is planned that allows higher level interactions to “confound” with single factors or other interactions. These are confounded or mixed in with other responses. If there are n factors, a half-fractional factorial will require 2^{n-1} runs at a minimum. For example, considering an experiment with five factors, one run at each factor would require 32 runs. A half-fractional factorial would cut this to 16. Consult a DOE subject matter expert (SME) for help with fractional factorial experiments.

There are several methodologies that utilize these basic experimental design types. Classical DOE was developed by Sir Ronald Fisher in England and promoted by Box, Hunter, and

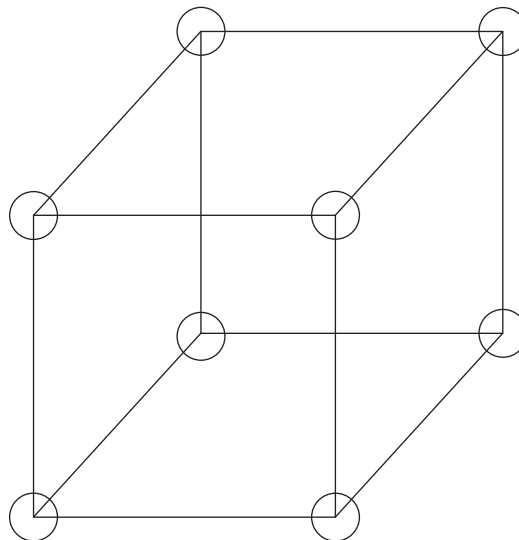


Figure 10 Full factorial.

Hunter in the United States starting in the 1970s. This type utilizes both full and fractional factorial designs. In the early 1960s, Dr. Genichi Taguchi began to promote, in Japan, a form of experimental design that uses a special set of fractional factorial designs. Although the forms Dr. Taguchi used were not unique, his approach generated a dramatic increase in DOE usage, especially among engineers. Dr. Taguchi made three major contributions to the field of DOE. First, he developed a DOE methodology that offered clearer guidance to engineers than earlier approaches. Second, he promoted the concept of robust design and demonstrated how DOE could be used to obtain it. Finally, he promoted the application of something he called the quality loss function. This could express in dollars how the enterprise and society in general are affected by variation from an optimal target.

Usually experiments are run at two levels. Occasionally, the engineer must experiment with factors at more than two levels. These may be attribute factors such as different materials or continuous variables such as temperature, pressure, and time. DOE handles all these, but the planning and analysis get a bit more complicated.

No matter what experimental design is chosen, it is very important to be aware of two key parts of an experiment. The first is randomization. Randomization means to plan the experiment carefully but run it in some type of random order. Using a random-number generator, picking numbers out of a hat, or any other method may accomplish this. The reason randomization is employed is to prevent some time-dependent factor from creeping into the experimental results. For example, a machine tool wears with use. If the experiment proceeds in a particular order with regard to the runs, the later runs will have the additional influence of tool wear. Randomization allows each factor-level setting combination an equal chance to experience a time-related factor. The other part that must be considered is replication. It is rare that an experiment is only run once at each factor-level setting combination. Even a full-factorial experiment is usually run with at least two replications so sufficient information is obtained for good analysis.

While this chapter has touched on the main types of experimental design, the author wishes to note that this has been a very rich field of research and innovation. As a result, several types of experimental designs were not covered but may be used for specific purposes, such as mixtures or the situations where output is nonlinear. These are discussed in the references provided for this chapter.

Regression Analysis. If process input parameter settings are recorded and can be correlated with output variables, regression analysis may show the relationship between inputs and changes in output. Start with simple linear regression. Use XY plots to see if there appears to be curvature in the input–output relationships. Curvature would suggest a more complex relationship with the variable, perhaps X^2 , X^3 , or an exponential for one or more of the factors (look back at Fig. 5). In all cases, go with the minimal formula and the lowest complexity that adequately expresses the relationship. Regression analysis can be explored using a Microsoft Excel spreadsheet or more sophisticated statistical packages. Data from DOE may also be used in regression analysis.

2.5 Improve/Innovate Phase in DMAIIC

The question for the improve/innovate phase in DMAIIC is “Using what we found in the analyze phase, what is the best way to close the gap between y_{current} and y_{desired} ?” If you’ve employed DOE in the analyze phase, you may already have the knowledge of what factors in the process need to change and what settings to use. If not, you might apply DOE here to discover. Alternatively, your team might explore additional factors or the factors in processes of key components.

Tools for Improve/Innovate Phase in DMAIIC

Cost/Benefit Analysis. Cost/benefit analysis can be as straightforward as its name or it might be more complicated in your particular case. If there is a payback, such as scrap avoidance

or improvement in production rate, the payback can be calculated as a function of time from the initial investment until the cost is recovered. If the improvement cannot be calculated financially, one can still assign a measure of goodness to the degree of improvement, but the resulting analysis is more subjective. The payback cost can be compared to the implementation costs of the proposed improvement(s). If there are options, the cost/benefit analysis can help with the decision.

Brainstorming. Brainstorming is fairly well known so little of this chapter is spent on it. Primarily, rules must be enforced to allow everyone on the team to participate. They must have an equal chance to participate without their ideas being rejected. If you are leading a brainstorming effort, take a look around and see if someone is not participating. Do not call them out for not participating! Rather ask them if they have some ideas they would like to share. Sometimes it takes a little encouragement. There are examples in business experience where the best ideas came from quiet process operators when they were finally encouraged to participate.

Risk Management. Risk management is not often used in process improvement except some regulated industries. Risk management adds a significant contribution to the effort. Properly tested out, the gain from your improvement should be reachable, but the implementation may have risks. These might include risks from changes to the process, changes to suppliers, and even changes to how users interface to the revised device. Assemble a multifunctional team to brainstorm all the potential risks to the change and determine ways they might be mitigated. A failure modes effects analysis (FMEA) can be used or a fault tree analysis (FTA). There are numerous examples on the Web of FMEAs. One example is at www.asourceofquality.com/resources.

Piloting the Solution. One way to manage risks of a process change is to pilot that change in a small, controlled manner. This allows you to see what issues surface and how these may be handled prior to turning the business over to the new process. It can be a career saver!

Validate the Improvement. After the improvement has been introduced into the process from a pilot or early test, it is advisable to repeat the process capability analysis or other data collection accomplished in the measure phase. If the process has been improved, changes in process capability and the resulting reductions in scrap or rework prediction are powerful statements for the team to use to explain the significance of their work and to obtain support in the implementation of the process improvement.

During the improvement phase, it is important to observe the effect of attempted improvements. As discussed in the previous phases, the team may have developed a plan for data collection. Data sampling and charts used during the measure phase may be repeated with the improvement inserted as a pilot trial. This can help validate the planned improvement. It is a well-known axiom that “correlation does not necessarily imply causation,” so be certain that the improvement results follow your process change. A good way to do that is demonstrate that you’ve “turned it on and off” by inserting improvements, running the process, set the process back, and rerun. The results should demonstrate a change if you have identified the correct factors.

Plan and Implement. The activity network diagram (AND), portrayed in Fig. 11, is a way for a team to schedule project tasks. The team can use simple cards or sticky notes to list project tasks. These can then be arranged in the anticipated flow (sequential, parallel, or combination) on a large wall with directional arrows indicating task relationships. The team can then estimate time to complete each task. The longest sequential path to complete for the whole project becomes the critical path. The graph also shows predecessor–successor relationships and the total task time can be calculated. This information can be an input to project management software. Microsoft Project or other project planning software may assist in the effort.

Documentation. As was discussed for Pareto analysis, the best use of resources demands that a team focus on the important items in the process. Critical to quality analysis or as it is

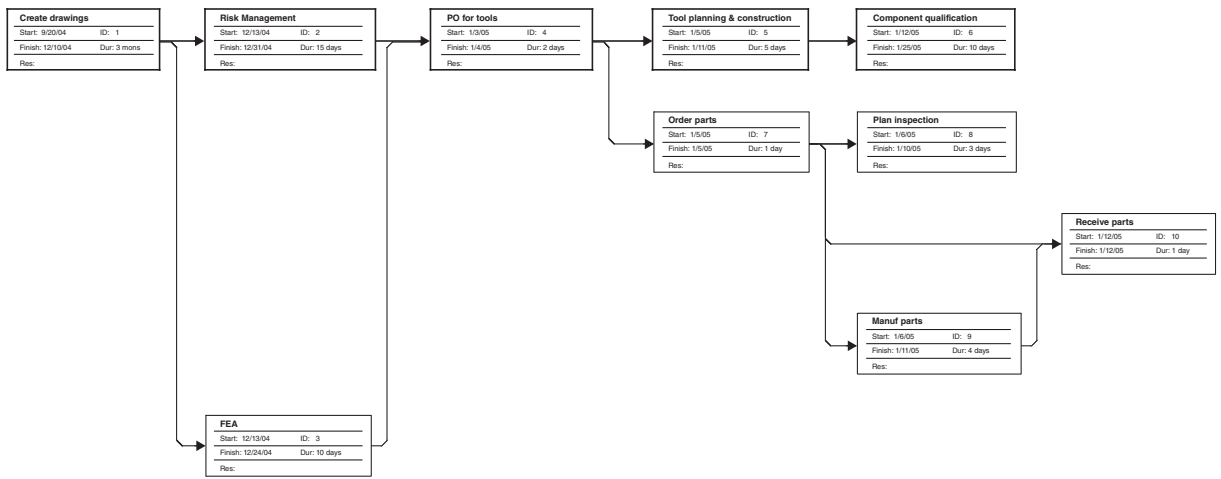


Figure 11 Network diagram.

often known, CTQ cascade has become known as the process to trace features of key customer importance into the process. This chapter will use CTQ cascade, as that is more descriptive. Now that you have spent a lot of time and resources understanding the process or product change, it is important to document what the team has learned.

In a CTQ cascade, a team takes the top critical to quality features for the output of the process and, through analyses or tests and experiments, relates them to process parameters or process inputs. For example, suppose a smooth paint finish on an auto body panel is a CTQ for our customers. From previous work, our team has found that critical painting process factors are spray pressure, paint mixing, and distance of the spray head from the body panel. Other process factors, such as temperature and time of application, are less critical for this. From this work, it is obvious that spray pressure, paint mixing, and distance of the spray head from the panel are critical process inputs to the CTQ. As time goes by, this linkage may not be so obvious to new workers and engineers on the process. To transfer this knowledge, we indicate the relationship using a CTQ cascade.

CTQ cascades often take the form of a tree diagram (see Fig. 12). This simple graphic shows the relationship very well. A process control plan is another tool that can demonstrate this relationship. A process control plan is a process work instruction generated in a word process document or spreadsheet. In this plan, it is convenient to show process settings in a tabular form. Linkages between a setting and a CTQ can be shown here. Cascades can also be shown in a spreadsheet. Early proponents of QFD often proposed using two or more QFD matrices to do this linkage. This is an excellent analysis approach but may be too difficult to maintain for a process work instruction. A process control plan fits that need very nicely.

Most teams that are new to this process will want to discuss what it means to be critical to quality. There are many things that are critical if left out or damaged. The way to look at CTQs is to determine what parts of the process are difficult to do or difficult to control. For example, process parameters that have tighter tolerances than normal might be CTQs. Another candidate for designation of CTQ is something that is new to the process. Continuing in the paint example used before, the addition of metal flake or pin stripes might be CTQs if they are not used in the normal process.

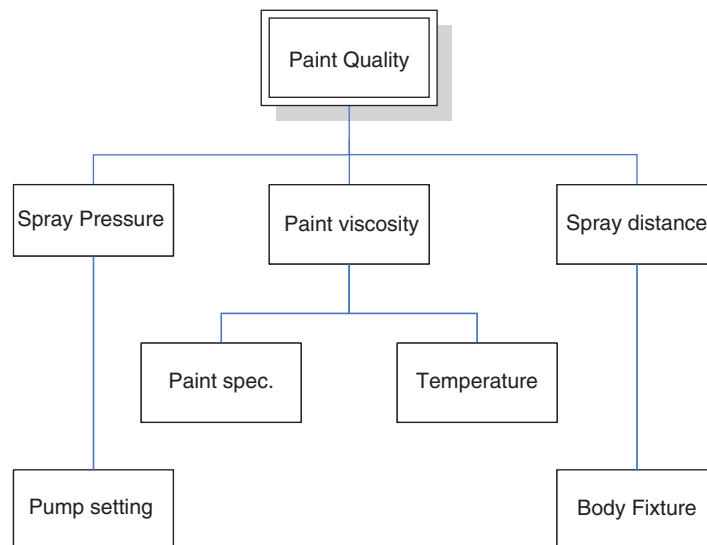


Figure 12 CTQ flowdown.

Change Management. Summarizing various definitions and descriptions of change management results in a description that change management is a way to help people, teams, and organizations shift from their current state to some future state. As a logical engineer, you may not see the need to help people with this shift. To you it might seem that the need for change is obvious and the change you propose is only logical! This may all be true, but you are closer to the issue than others. It is up to you to help them prepare for and make the change necessary to capture the improvement.

Change management is *mostly* about communication. Go back to your stakeholder analysis. Look at the various stakeholders and see what you need to do to move them to the level necessary to be supportive of the change. You did do a stakeholder analysis, didn't you?

2.6 Control Phase in DMAIIC

The major question for the control phase is, "How do we ensure that the improvements remain after the team moves on to other things?" Now that you have found the important values for the critical process inputs, it is important to make them *stick*.

Tools for Control Phase in DMAIIC

Control Plan. A control plan is a work instruction for a process. Control plans can take any form, but they are usually maintained as a word processing document or spreadsheet under revision control. They can take the form of a word document with complete process instructions or a table of process parameters with their settings and ranges. As stated in the discussion of CTQ analysis, factors in the process that have a key effect on CTQs can be identified. Key items to cover in a control plan are:

- Process step or phase
- Order of actions (if sequence is critical)
- CTQ linkage (if any)
- Target setting
- Allowable range
- Calibrations needed
 - Sampling plan (number of samples, what to measure, what measurement tool to use, and how often to sample)
- MSA for measurement tools
- Reaction plan (orderly shutdown)
- Safety measures and equipment

Evaluating Results. If you have not performed a pilot or done a validation of the improvement in a previous phase, it should be done in the control phase. Even if you have performed a pilot or validation, make sure data are available from your control plan to evaluate the results against the target value.

Monitoring. Using the control plan, put in place data collection and reporting that keeps watch on the measures you improved and any other measures required for this process.

Statistical Process Control (SPC) or Control Charts. Control charts in the control phase seem a natural fit, and they are. SPC is a technical tool that came into general use early in Japan. After less than satisfactory first attempts at deployment of SPC, many companies are finding it to be useful for reducing defects, lowering defect rates, and making business key processes more consistent and dependable. The key to successful use of this tool is to understand what SPC does and does not do.

SPC is the application of a statistical method, usually in a graphical form. It is used to detect when a process *may* have been influenced by a “special cause” of variation. Dr. Walter Shewhart, who developed the earliest concepts and applications of SPC, divided process variation into two types. One he described as “common cause” or “normal” variation in his writings. Common-cause variation comes from the many factors in the process varying and interacting with each other. For example, in a drill process, there is drill splay (wobbling of the drill bit around its axis), variation in bits, and variation in material hardness, etc. These interact and result in a variation of hole size, position, and degree of roundness of the hole. The second form of variation described by Dr. Shewhart is referred to as “special cause” variation. Continuing with the drill example, insertion of the wrong bit size would result in a change of the hole size. This shift of hole size is not “normal” but can be *assigned* to a process error. Other examples of special causes might be untrained operators, improperly maintained machines which exhibit more variation, changes in material, changes in bit manufacturing, and changes in material clamping technique, among others. Figure 13 shows one of the first and most used charts, the \bar{X} -bar and R chart. This is also noted as \bar{X} and R in mathematical nomenclature, where \bar{X} stands for the subgroup average and R stands for the subgroup range. A subgroup is a sample that is taken periodically in the process. In the example, a point is out of tolerance in each chart. A subgroup usually consists of a sample of 2–10 units for this type of chart. This type of chart can detect a shift in the sample average (through the \bar{X} portion) and the sample standard deviation (through the R portion of the chart). Together, these two charts signal changes of the process average and process variation of one variable measure.

The reason to make such a distinction between these two sources of variation is to separate the *manageable* from the *unmanageable*. Special causes of variation can be identified

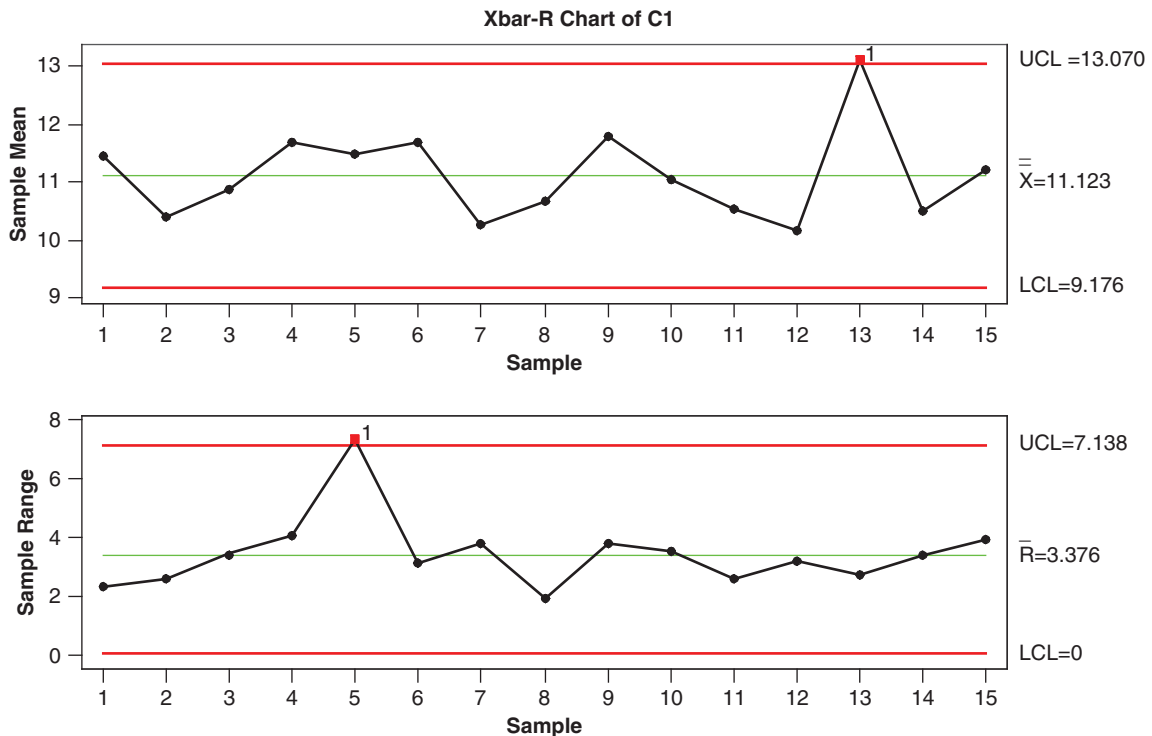


Figure 13 \bar{X} -Bar and R .

and removed or prevented from entering the process again. Often, these changes may be easily made. Normal causes of variation usually may be removed or reduced only by changes to the process. Of course, this involves management commitment and possibly capital expenditure to change the process. Continuing on the drilling example, higher accuracy and repeatability might require a process change to laser drilling or a water jet. Obviously, this requires different machinery and often such a change is not trivial.

How does SPC fit into this? Dr. Shewhart, working in an AT&T Western Electric plant, noticed that their processes had excessive variation and the operators were constantly adjusting the process. He suspected they were adding to the process variation by these adjustments. He sought a way to determine when a process adjustment was necessary and proposed the use of SPC and SPC charts as a way to *signal* when a process may have been influenced by a special cause of variation. When the signal occurred, operators, engineers, and management could pursue adjustments or investigations as seemed appropriate.

SPC charts come in many forms, but in general, all plot one or more statistics (a descriptive measurement from a unit or sample) on a form of line chart. The line chart also contains warning limits and control limits depending on the chart type. Look back at the example in Fig. 13. The control limits are derived from past stable process data and usually represent some long-range average of the measurement \pm three standard deviations for the statistic. For statistical reasons, some charts do not have a lower control limit. Most of the measurements used for SPC follow a normal distribution (helped by a statistical phenomenon called the central limit theorem). This means that follow-on measurements from a process that is not affected by special cause variation will stay within control limits 99.73% of the time. Reversing this logic, a point outside of a control limit would only happen 0.27% of the time when no special cause is present, so when such an event happens, it is most likely the result of a special cause of variation. Process investigation should be employed to find and remove the special cause.

In addition to watching for points outside of the control limits, SPC charts may send other signals. SPC practitioners apply tests for *patterns* that signal the effect of a special cause variation. For example, a pattern of seven points in a row, increasing or decreasing, is a pattern that rarely shows naturally. Such a pattern indicates the likely presence of a special cause of variation, even if no control limit has been breached! The following are some rules for abnormal patterns in SPC charts:

- One point beyond a control limit
- A run of seven or more points either up or down or consecutive above or below the centerline
- Two of three consecutive points outside 2σ but still inside the 3σ line
- Four of five consecutive points beyond 1σ

While SPC deals with in-process measures, often our only significant way to measure the process result is by measuring the performance of the finished product. For example, when we assemble an electronic circuit, there are measurements that can be taken in the process, but the final circuit performance can only be measured by a final functional test. As with in-process measures, final performance variation is a function of normal and special cause variation. SPC can also be used in the case of final process performance to determine if an investigation of special cause variation is warranted. This is often referred to as *statistical quality control (SQC)* to differentiate it from process control. The same theory is used, but the charts are sometimes slightly different as different statistics are employed. We should note that SQC should not be used as a substitute for SPC. Since SPC works with in-process measures rather than the end of the process, it offers faster detection and correction of problems.

SPC and SQC are powerful tools, but they essentially do only one thing—they identify when a process has been influenced by something not usually a part of the process. When that

occurs, process engineers and operators can investigate for the cause and remove or prevent it, returning the process to what is the normal state. This is accomplished by examining the control plans and documentation accumulated from the last DMAIC on the process. If it is not clear what has affected the process, a new DMAIC action may be warranted to put the process back on track.

Closure. Formally hand off the improved process to the process owner if it is someone else. Be sure to reward the members of your team. It helps the overall process to do it as publicly as your culture supports.

3 DMADV

As noted earlier in the chapter, DMADV will be used if you are developing a new product or process. DMAIC will be used if you are improving a product or process, even if the process is rather ad hoc. If you are working on something and nothing exactly like it existed before, you probably need DMADV. If you are radically redesigning a product or process, you probably need to use DMADV.

3.1 Define Phase in DMADV

The first step is fairly simple but exceedingly important, just as it was in the DMAIC flow. The question that needs to be answered here is, “What is it that I or my team needs to do for the business or our customers?” In the define phase the team will detail what needs to be accomplished and it is especially important that there be a written agreement between the team and management on the required changes, timeline, and budget of the project. Some projects can span a significant amount of calendar time, and memories of the original agreement can differ when the team is ready to wrap the project up.

Tools for Define Phase in DMADV

Charter. Like the charter in DMAIC, a formal charter defines the relationship of your project to the needs of the business. It is important to work out what you will deliver, your project budget, and expected timelines so that marketing efforts may be prepared. It is also important to define who in the organizational leadership is sponsoring the project. It will be necessary to work through decisions and risks as the project moves along. Sponsors are the go-to people to clear the way for the project. An example charter can be seen at www.asourceofquality.com/resources.

Multigenerational Planning. Usually, projects for goods and sometimes for services have more than one generation planned for the pipeline. It is also possible that they do not have this planned, but they should. If follow-on generations are defined, it is important to be thinking about the next generation as the current project moves through phases. This prevents current decisions from closing the door to a future version. It also happens that some members of the organization get so excited about the current project that they want to put everything possible into it. That might not be a good business strategy. It may not be wise to let the current product or process include all possible future options.

In DMAIC we talked about an in-scope/out-of-scope tool to keep track of those things that come up for improvement as the project continues. A multigeneration plan (MGP) serves a similar function for DMADV. Good ideas may come up during a project, but their inclusion would be too extensive for the current scope. An MGP serves as a way to avoid losing those ideas.

Project Plan. Project planning is a large topic that is covered well in other writings. It will not be addressed here other than to state that you must have a plan to be successful.

SIPOC. See the DMAIC discussion on SIPOC. For a new product, make sure you define your customers, including the different levels and types of customer. For example, a surgical device will obviously have the surgeon as a customer, but the patient, nurses, central sterilization, and hospital administrators are also customers. Understand what each level of customers expects of your product. Do not be surprised if there is some tension between stakeholders. For example, a surgeon may want a new surgical device, but the hospital and insurance providers may not be as willing to pay for a new technology. This can be especially true if the benefit is not established.

Risk Register. In the risk register, risk will be defined as business or more specifically project risk. Keep a risk register and discuss it with your sponsors. It should contain identified risks, a simple measure of criticality (H-M-L), impact (schedule, cost, etc.), who on the team is handling it, and what are possible mitigations. It must have a way to indicate when a risk is closed or if it is still open. It is important for your sponsors to know what risks you are managing and what the mitigations might be. They should not be surprised when a risk emerges.

3.2 Measure Phase in DMADV

In the measure phase for DMADV, we are going about defining the service or product. The thinking here is to understand what your customer(s) need and to then develop requirements for your goods or services. Just as in the DMAIC process, the rest of the flow will be about determining $y = f(x)$. In the measure phase, you will need to understand what values of y are important for the customers, so the question here is, “What is important to my customers?”

Tools for Measure Phase in DMADV

Voice of the Customer (VOC). Just like it sounds, we are going to obtain and refine statements of what the customer wants, preferably in his or her own voice. Original VOC should be the customers’ statements in their words. The team will create a flowdown of requirements from VOC using QFD or CTQ analysis.

Kano Model. What is now called the Kano model was expressed by Dr. Noriaki Kano in 1984. He expressed the concept that customers had expectations of a product or service. Some of the expectations were basic needs. If absent, these would render the product or service unacceptable. There are attributes that can be described as “delighters.” These bring more of a sense of excitement or delight to customers. It is also known that the delighters of past products and services can quickly become expectations today. See Fig. 14 for a graphical example of the Kano model.

Quality Function Deployment. QFD is an exceedingly important tool that might be overlooked because its purpose is misunderstood or from fear of its complexity. At its heart, QFD is a form of requirements flowdown in graphical form. Engineers understand requirements flowdown. Nonengineers will benefit from the graphical form of QFD. After determining important features and needs in the product from the voice of the customer, the best tool to use for capturing these and relating them to the design elements is QFD. You will recognize the core form of QFD as a simple L-shaped matrix. QFD was initially applied in the 1960s in the Kobe shipyards of Mitsubishi Heavy Industries of Japan. It was refined through other Japanese industries in the 1970s. Dr. Donald Clausing first recognized QFD as an important tool. It was translated into English and introduced to the United States in the 1980s. Following publication of the first book, *Better Designs in Half the Time*, it has been applied in many diverse U.S. industries.⁵

At the heart of applying QFD are one or more matrices. These matrices are the key to QFD’s ability to link customer requirements (referred to as the voice of the customer or customer *WHATs* in QFD literature) with the organization’s plans, product or service features, options, and analysis (referred to as *HOWs*). The first matrix used in a major application of

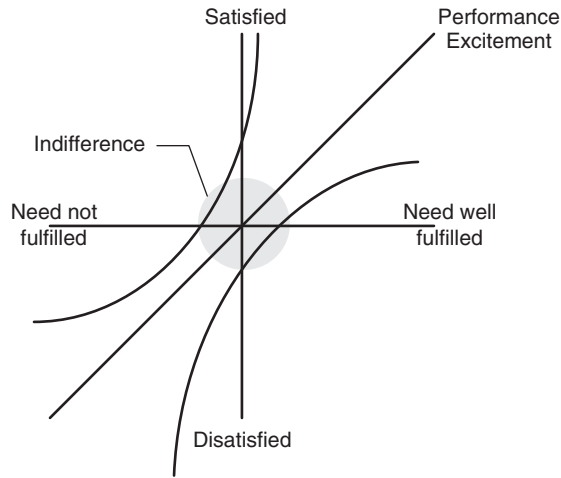


Figure 14 Kano model.

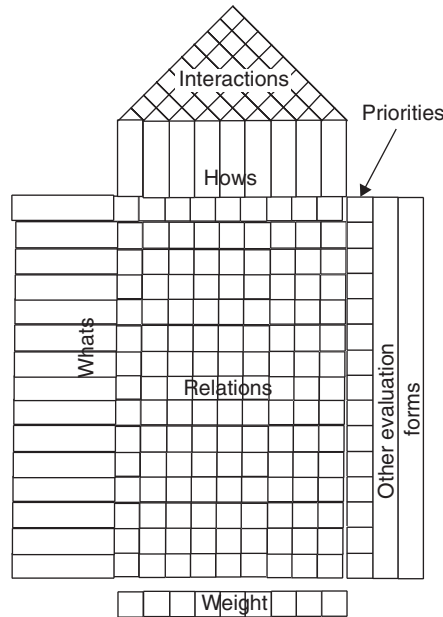


Figure 15 House of Quality.

QFD will usually be a form of the A-I matrix (Ref. 5, pp. 2–6). This matrix often includes features not always applied in the other matrices. As a result, it often takes a characteristic form and is called the House of Quality (HOQ) in QFD literature. Figure 15 presents the basic form of the HOQ.

The A-I matrix starts with either raw (verbatim) or restated customer WHATs and along with corresponding priorities for the WHATs. Restated customer WHATs are generally still qualitative statements, but with more specificity. For example, if the original VOC was for the car dash to have a cup holder, the restated WHAT might be that it has room for a 16-ounce cup

of coffee. The priorities are usually coded from 10 to 1, with 10 representing the most important item(s) and 1 representing the least important. These WHATs and their priorities are listed as row headings down the left side of the matrix. Frequently we find that customer WHATs are qualitative requirements that are difficult to directly relate to design requirements, so the project team will develop a list of substitute quality characteristics and place these as column headings on this matrix. The column headings in QFD matrices are referred to as HOWs in the QFD literature. Substitute quality characteristics are usually quantifiable measures that function as high-level product or process design targets and metrics.

The term “substitute quality characteristic” may appear ambiguous. The best way to think of this is to consider the fact that verbatim customer requirements may be stated in words that cannot be directly translated to equations. For example, when a user describes the need to make a kitchen appliance “easy to use,” there is no way to put that as a specification and measure the output. However, it is still the *voice of the customer*. So, using QFD, it would be placed down the left column as a WHAT. The team would then place ways to make the device easy to use along the top as column headings. Examples of ways to achieve “easy to use” might be well-marked controls, no more than one knob, easy to turn, easy cleaning, or automatic sensing of the appropriate setting. Each of these becomes a HOW and, if it relates to the WHAT, becomes a substitute quality characteristic for the easy-to-use WHAT.

The relationships between WHATs and HOWs are identified using symbols such as ● for high relation, ○ for medium relation, ▼ for low relation, and *blank* for no relation.* These are entered at the row/column intersection of the matrix. The convention is to assign nine points for a high relationship between a WHAT and a HOW, with 3, 1, and 0 for medium, low, and no correlations, respectively. The assignment of points to the various relationship levels and the prioritization of customer WHATs are used to develop a weighted list of HOWs. The relationship values (9, 3, 1, and 0) are multiplied by the WHATs priority values and summed over each HOW column. These column summations indicate the relative importance of the substitute quality characteristics and their strength of linkage to the customer requirements.

The other major element of the A-I matrix is the characteristic triangular top (an isosceles triangle) which contains the interrelationship assessments of the HOWs. This additional triangle looks like a roof and gives the QFD matrix the profile of a house, hence its nickname, the House of Quality. Look back again at Fig. 15. The roof contains indicators that show the relationship between “HOWs.” The best way to think of this is to consider what would happen to the other design elements if each one is increased in turn. Consider, for example, a QFD for a car. In response to customer needs and wants, we intend high mileage and ease of operation. To achieve high mileage, we also intend to forego power steering and automatic transmission. The latter decision would improve mileage, but it would have a detrimental effect on ease of operation for most drivers. The relationship between HOWs is noted in the “roof” by four symbols, ++ (strong positive relation), + (positive relation), – (negative relation), – – (strong negative relation), or blank (no relation). The positive relationships indicate that increasing the design attribute (HOW) will cause a corresponding increase in the connected HOW. There is no numeric analysis done with these relationships. These are informative for potential trade studies.

Other features that may be added to the A-I matrix include target values, competitive assessments, risk assessments, and others. These are typically entered as separate rows or columns on the bottom or right side of the A-I matrix.

The key output of the A-I matrix is a prioritized list of substitute quality characteristics. This list may be used as the inputs (WHATs) to other matrices. For example, in Fig. 16, we

* The symbols shown are commonly used in QFD. For more information about QFD see <http://qfdcapture.com/default.asp>. QFDcapture is a leading software tool for QFD.

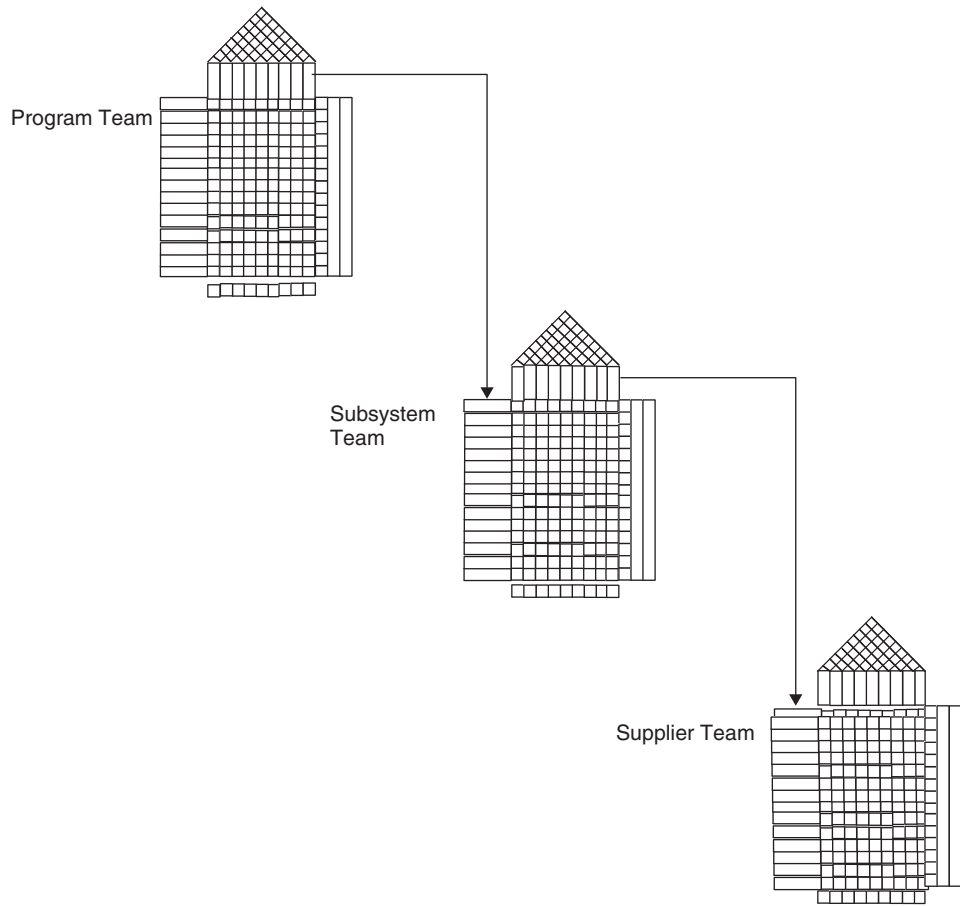


Figure 16 Matrix flowdown.

show the HOWs of the program team feeding requirements (WHATs) to the subsystem team and the subsystem team HOWs feeding requirements (WHATs) to the suppliers.⁵

Design Scorecard. A design scorecard is used in the DMADV process to communicate to the sponsors how well the team is doing on achieving customer needs, especially the CTQs. On the left side of the table, put customer requirements. In adjoining columns place target values and how they will be measured. As the project evolves, you will fill out the actual values achieved by the team.

If all goes well, you will achieve all the customer requirements. In some cases, one or more will not be completely achieved or only partially achieved. The design scorecard will enable your team to communicate this to sponsors and the organization so they can evaluate the impact of the design performance. See www.asourceofquality.com/Resources for an example of a design scorecard.

3.3 Analyze Phase in DMADV

The analyze phase is the place where the team will investigate possible conceptual designs and begin the design effort. This will also be the place where different design concepts might

be subjected to trade-off analysis. In the $y = f(\mathbf{x})$ form, the analyze phase will begin the first exploration for the important values in the set of \mathbf{x} . These will be top-level requirements. The question in the analyze phase is, “What are the requirements and their specifications that enable the voice of the customer?”

Tools for Analyze Phase in DMADV

Concept Generation and Evaluation. Concept generation can take several forms at different companies. You may do simulations, mockups, brainstorming, or sketches.

High-Level Process/Product Design. Using QFD (started in the measure phase), begin to flow down customer requirements into the system requirements and subsystem requirements for the concept that has been selected. System requirements are those specifications that directly translate from VOC.

Simulation and Modeling. Where possible, leverage simulation and modeling to get feedback. It is useful even to employ mock-ups that are not functional. Money spent in this phase has greater leverage than if the design is allowed to go to completion and you miss the customers’ expectations. You may do an early device with short-run tooling or 3D printing. It might be fairly crude, a form of “Franken-design.” The important thing is to get feedback from the customer.

Design of Experiments. In the analyze phase, there may not be much available for experimentation, but the team can begin to plan those experiments that will be needed in the design phase.

Reliability Assessment. Even though the design phase has not been started, it is possible to begin to assess the reliability of known subsystems. It is also the time to begin planning how reliability will be assessed for the new product or service. If you did not do it in the measure phase, be sure you capture customers’ expectations on products.

Risk Management. Begin fault trees or design failure modes and effects analysis to identify the mitigations necessary in development in the design phase. Perhaps some hazards will require specific design elements to bring risks into an acceptable level. Now would be the time to plan for this.

3.4 Design Phase in DMADV

The design phase is where the team will develop the detailed design, process, and service requirements for the new product. Critical requirements will be flowed down into the detailed elements of the design, such as subsystems, software, and components. The members of the team charged with developing the manufacturing processes will specify and construct the manufacturing stations to include the fixtures, assembly equipment, and steps for personnel to follow. The team will prepare for verification and validation tests. The question for this phase is, “What are the important details in components, subassemblies, and processes that have to be right to achieve customer expectations [$y = f(\mathbf{x})$]?”

Tools for Design Phase in DMADV

Detailed Design Elements. Now that the system has been specified, the detailed elements below the system level must be specified. This includes the subsystems that are purchased as well as individual components and software. QFD can be a great tool here as requirements flowdown may be continued to these elements of the system. This will maintain the trace from VOC to the key design elements of the system.

Some companies are using relational requirements databases to flow requirements down. A relational requirements database links design elements in a manner similar to the graphical framework of QFD, but the graphical relationship is more difficult to display and QFD lends

itself to development of requirements and specifications that are not directly part of the design. These include service elements and manufacturing support. It can also identify sources of risk.

Design for X (Environment, Human Factors, Manufacturability, etc.). Over the last few years additional “voices” have been added to the design world. Designing for the environment and sustainability are big at the present time. Designing for human factors is important, especially in the areas of high safety requirement such as transportation, system control, and medical devices.

Designing for manufacturability and assembly has been important for some time. The techniques that are involved are reasonably simple (see reference materials at the end of the chapter). They involve looking at the design from the perspective of assembly personnel and making their efforts easier and faster. This means:

- Minimizing the number of components (fewer components mean fewer errors)
- Minimizing the number of different fasteners (fewer types of fasteners mean fewer items to stock and fewer tools needed for assembly)
- Designing so the device only goes together one way
- Utilizing gravity and special fixtures to hold main assemblies in place. Make it so the assembler does not have to access connections in a difficult position, such as upside down.

DOE and Statistical Tolerancing. Design of experiments in the design phase can be used to “tune” in the process for the design and to reduce variation. In the design phase, DOEs are focused more on finding and controlling sources of variation and setting the limits on processes. Output may also be used to develop component tolerances through the use of statistical tolerancing.

Statistical tolerancing is using the voice of the process to develop the tolerances for the features on parts so that the part specifications can be met by the process. For example, if the long-run variation in a process is ± 0.05 in., it will be costly and frustrating to specify that a part must meet ± 0.03 in. If the part must truly meet that specification, the existing process must be improved or another process must be used to create the part feature. Statistical tolerancing can be developed using the formula $SpecLimits = \bar{X} \pm k * s$ where \bar{X} represents the process average and s is the process standard deviation. The value k represents a constant which contains a factor for the percent tolerance and the confidence value adjusted for the sample size, where \bar{X} and s are developed. Statistical tolerancing may also be developed from process capability studies. If the process exhibits the behavior that it does not follow a normal distribution, additional tolerancing may be required. See additional reference materials at the end of this chapter.

Some companies specify part feature tolerancing through stackups, which analyze the potential part interference at key areas of interaction. There is nothing wrong with using stackups, but if the process cannot create the feature within the tolerance, the stackup is misleading. It might be possible to use the stackup to determine where additional tolerancing may be “taken” from one part and applied to another, assuming the result is supported by the processes and results in acceptable performance.

Process Simulation. A current technique for managing very lean processes is to simulate the flow of goods through the process. This can identify potential bottlenecks and the steps in the process that will pace the flow of material. Simulation can be as extensive as desired (and as money allows), but the most useful simulations need not be graphically intensive. A useful simulation would simulate the average time a unit stays in each operation (including some variation), movement time to the next step, and anticipated scrap/rework rates. Such simulations can identify the average throughput rate and highlight the operations that will pace flow. Each operation that paces the flow is a process where the overall availability must be maximized. This

might be termed a “bottleneck” process or a “pacing” process. This means that this process must be on preventive maintenance and scheduling must be designed to keep this process busy.

Verification and Validation Plans. As soon as the major design requirements are firmed up, the team must plan the way to verify and validate them. See the next section for an explanation of the difference between verification and validation. It is especially critical that quality and regulatory design requirements be verified and actual users should validate that the design hits the mark.

Control Plans. The control plan, as its name implies, is a plan for controlling the final quality of the product or process. It includes defining how, in the process, the quality of design elements will be controlled to assure they are correct. Control plans should include:

- Incoming material inspection/certification
- Component and subsystem inspections/tests
- Assembly inspections/tests
- Final device acceptance testing

3.5 Verification and Validation Phase in DMADV

The verification part of this phase is where the team will assure that the finished design meets the engineering requirements and the process/product performs as intended. The validation part of this phase is where the team makes sure that the completed product/process meets the needs of the customer. The difference between verification and validation is not only the customer involvement in the latter but also the fact that validation checks to see if the customer VOC was adequately translated into engineering requirements. Verification is typically done by the team or other engineers/operators. There are two questions for the verification and validation phases. The first question is, “Have we produced the new product or process to our specifications?” This can be restated as, “Have we hit the important values of x that were determined in the previous phases?” The second question is, “Does the output product or process meet the users’ needs?” The last question assesses the full process of translating customer input (VOC) into engineering and producing an appropriate output.

Tools for Verification and Validation Phase in DMADV

Verification Tests. Verification tests will be done with documentation and planning appropriate for your organization’s expectations. It is *always* important to define the criteria for success ahead of time to avoid the appearance that the results were taken and then the criteria were set.

Talk to statistical personnel to define good statistical criteria for your tests. Some tend to define simple tests with a small sample size, only to be surprised that the long-run results do not match the “test results.” A well-formed test incorporates the following concepts and the test designer understands their potential impact:

- Proposed or null hypothesis
- Alternate hypothesis
- Type I error and its probability (called the alpha value)
- Type II error and its probability (called the beta value)

Type I error is the potential that the test will reject the null hypothesis when it is actually true. Type II error is the error that the test will accept the null hypothesis when the alternate hypothesis is actually true. It is important to understand that neither of these probabilities ever goes to zero, but it is possible to structure a test that makes them acceptable. Many companies

use a target alpha value of 0.05 or less, depending on the risk. The beta value may be set at 0.10 or less, once again depending on the risk.

Validation Tests. As noted previously, validation is only truly done by your customers. Some may try to use internal marketing or sales as a stand-in for customers, but it is truly only the customer who has the final say on whether you have translated VOC into a product that fulfills. There is often a desire to short-cut validation in order to “get to the market” with the idea that actual sales are the true validation. This is true up to a point, but taking the unit to market entails a large commitment on the part of the organization. Be sure to include the product instructions in your validation tests.

Process Documentation. Do not forget to update the process documentation.

Transfer. Depending on your organization, development of a new service or project may involve transferring the result to a group responsible for the on-going maintenance and further deployment of the new product or service. Follow your internal transfer practices. If your organization does not have any, make sure your documentation is in order and the new management team is trained in the important findings from your development. Especially pay attention to the items that were directly related to key VOC and critical to quality.

Key Learnings. Whether your organization does this normally or not, schedule some time to transfer the key learnings from this project to others in the organization that could benefit.

4 SUMMARY

Two different paths of TQM have been presented for the mechanical engineer to be successful at leading projects of design or improvement. Numerous tools are available at each step of the way. Questions were provided to guide the appropriate tool selection and to understand where the team is in the process. Mechanical engineers are encouraged to continue practicing the processes of TQM for their businesses and personal careers.

REFERENCES

1. QP Staff, *One Size Fits All, Quality Progress*, ASQ Press, Milwaukee, WI, April 2012.
2. S. S. Chakravorty, “Where Process-Improvement Projects Go Wrong,” *Wall Street Journal*, June 14, 2012, available: <http://online.wsj.com/article/SB10001424052748703298004574457471313938130.html>, accessed October 21, 2012.
3. S. Covey, *The Seven Habits of Highly Effective People*, Simon and Schuster, New York, 1989.
4. Automotive Industry Action Group (AIAG), *MSA-3: Measurement Systems Analysis*, 3rd ed., AIAG, Cincinnati, OH, 2002.
5. B. King, *Better Designs in Half the Time*, Goal/QPC, Salem, NH, 1989.

REFERENCES NOT CITED

The dynamic nature of the World Wide Web means that some of the site references below will change so that the links may not work directly. Usually, the information is maintained by the organization hosting the site, but it might be moved to a different or renamed page. If the direct site does not work, go to the home page of the organization or use a search engine to reacquire its location.

American Society for Quality (ASQ): ASQ has always been a source for quality-related information. The society and its members have been active in all quality-related initiatives. Resources, training, and references are available at <http://www.asq.org>.

National Institute for Standards and Technology (NIST): NIST offers information on the Baldrige National Quality Award and also offers guidance with measurement systems. Besides information on the Baldrige

National Quality Award, NIST maintains an on-line *Engineering Statistical Handbook* in conjunction with SEMTECH. This may be accessed at <http://www.itl.nist.gov/div898/handbook/index.htm>. Several specific references in this chapter use pages in this handbook.

RECOMMENDED FURTHER READING

- Design for Lean Six Sigma: A Holistic Approach to Design and Innovation*, R. Jugulum and P. Samuel, 2008, John Wiley & Sons, Hoboken, NJ.
- Engineering Statistics*, 5th ed., D. C. Montgomery, G. C. Runger, and N. F. Hubele, 2010, John Wiley & Sons, New York.
- Environmentally Conscious Manufacturing*, M. Kutz (Ed.), 2007, John Wiley & Sons, New York.
- Environmentally Conscious Mechanical Design*, M. Kutz (Ed.), 2007, John Wiley & Sons, New York.
- Practical Reliability Engineering*, 5th ed., P. O'Connor and A. Kleyner, 2012, John Wiley & Sons, New York.
- Practitioner's Guide to Statistics and Lean Six Sigma for Process Improvements*, M. J. Harry, P. S. Mann, O. C. De Hodgins, R. L. Hulbert, and C. J. Lacke, 2010, John Wiley & Sons, New York.
- The QFD Handbook*, J. B. ReVelle, J. W. Moran, and C. A. Cox, 1998, John Wiley & Sons, New York.
- The website www.weibull.com offers some powerful software for analysis of reliability data and for planning reliability engineering effort. There are also great resources in the form of online e-books for reliability.
- Various statistical resources: Normal probability applet, available: <http://www.rossmanchance.com/applets/NormalCalcs/NormalCalculations.html>, accessed September 16, 2012.

CHAPTER 23

REGISTRATIONS, CERTIFICATIONS, AND AWARDS

Cynthia M. Sabelhaus and Eric H. Stapp
Raytheon Missile Systems Company
Tucson, Arizona

1 INTRODUCTION	667	2.8 Mission Assurance—Missile Defense Agency Assurance Provisions	674
2 REGISTRATION, CERTIFICATION, AND ACCREDITATION	668	2.9 Capability Maturity Model Integration	675
2.1 National and International Standards	669	2.10 ISO 14000: Environmental Management System Requirements/EMAS	675
2.2 ISO 9001: Quality Management System Requirements	669	2.11 ISO 22000: Food Safety Management Systems—Requirements for Any Organization in the Food Chain/HACCP	677
2.3 ISO 9001: Certification/Registration	670	3 QUALITY AND PERFORMANCE EXCELLENCE AWARDS	678
2.4 AS 9100: Quality Management Systems—Requirements for Aviation, Space and Defense Organizations	670	3.1 Deming Prize	678
2.5 ISO/TS 16949: Quality Management Systems—Particular Requirements for the Application of ISO 9001:2008 for Automotive Production and Relevant Service Part Organizations	672	3.2 Baldrige National Quality Award	680
2.6 ISO 13485: Medical Devices: Quality Management Systems—Requirements for Regulatory Purposes	673	3.3 U.S. State Quality Awards	684
2.7 TL 9000: Telecom Quality Management System	674	3.4 Shingo Prize for Operational Excellence	684
		3.5 Quality Awards around the World	685
		3.6 Industry-Specific Quality Awards	688
		3.7 How Do They Compare?	688
		REFERENCES	688

1 INTRODUCTION

In recent decades, the concept of quality has evolved from inspection of items during or after production, to development and control of processes, to optimizing the entire quality management system. The International Organization for Standardization (ISO) defines this as a “management system to direct and control an organization with regard to quality as formally expressed by top management.” In effect, the quality management system encompasses much of what an organization does to provide customer satisfaction.

Early attempts to define and document quality requirements include MIL-Q-9858, the U.S. Military Specification on “Quality Program Requirements.” This document outlined general requirements for maintaining product quality but was generally weak in the area of management’s role. As the total quality management (TQM) movement took hold in the 1980s, more emphasis was placed on the role of top management to proactively drive the quality management system. Beginning in 1987, the Baldrige National Quality Award (BNQA) further emphasized management’s role in quality.

As various government and commercial entities tried to define quality requirements, a multitude of similar documents arose with varying requirements from one industry or country to the next. This created difficulties for companies that supplied multiple industries—the rules and requirements changed for each customer they served.

An attempt to rectify this situation resulted in the development of the ISO 9000 series of documents defining standards for quality management systems. However, in subsequent years, various industries have developed their own standards, often including the ISO requirements in their entirety, but with new requirements added. The U.S. Department of Defense (DoD) has also had a major role in reshaping the management aspects of quality standards. After years of acquisition reform and a move toward simplifying requirements to prime contractors and their supply chain, the DoD spurred the creation of a number of detailed requirements, including the Capability Maturity Model Integration (CMMI®) and the mission assurance provisions (MAPs). As these requirements were flowed from prime government contractors to their subcontractors and suppliers, the cascading effect has had a major impact on the management of business throughout the world. For many suppliers of generic products that serve several industries, the complexities of multiple standards and varying certification requirements add to the level of difficulty in doing business.

As companies engage in the process of achieving certifications, registrations, and awards, mechanical engineers will undergo audits and site visits testing their understanding of applicable quality management system requirements that range from configuration control to the manner in which parts are procured. Some engineers will be asked to help their companies create the processes and/or documentation required to achieve certification or apply for an award. This chapter provides a general overview of the most widely recognized programs. Keep in mind that standards are revised periodically, and award criteria may be updated annually. Use the Web links provided to obtain the latest information.

2 REGISTRATION, CERTIFICATION, AND ACCREDITATION

While the concept of certifying or registering quality systems to an industry or international standard has become the norm throughout the world, the terminology is often misunderstood. For all practical purposes, the terms *certification* and *registration* are interchangeable. When a company seeks validation of its ISO or industry-specific quality management system by hiring a third-party registrar, the quality system is certified as meeting the requirements, and the registrar issues a certificate. The certification is then entered in a register of certified companies. Thus, companies meeting the requirements of the standard are both certified and registered. The term *certification* is most often used for this process in Europe. In the United States, it is more common to hear the process called *registration*.

The terms *certification* and *registration* should not be confused with *accreditation*, which is the procedure by which an authoritative body gives formal recognition that a body or person is competent to carry out specific tasks. Accreditation bodies have been set up in a number of countries to evaluate the competence of certification bodies. An accreditation body will accredit, i.e., approve, a conformity assessment body such as a registrar. The registrar is then approved to review organizations for compliance to specific standards such as ISO 9001 or ISO

14001 and recommend that organizations be certified as meeting those requirements. For ISO standards, there are over 80 accreditation bodies throughout the world, including the American National Standards Institute (ANSI)/ASQ National Accreditation Board (ANAB) in the United States, the United Kingdom Accreditation Service (UKAS), and the Standards Council of Canada (SCC). Many of these bodies have reciprocal agreements to recognize each other's accreditations.¹

2.1 National and International Standards

Technical standards have played a large part in enabling the creation of a global economy. At the same time, they have enhanced living standards by ensuring common standards for safety. The first efforts in international standardization began with the International Electrotechnical Commission (IEC) in 1906 and the International Federation of the National Standardizing Associations (ISA) in 1926. By 1942, the efforts of both organizations were abandoned.²

The need for expansion of the standardization efforts was identified as countries worked together during World War II. In 1946, delegates from 25 countries met in London and decided to create a new international organization to facilitate the international coordination and unification of industrial standards. The new organization, the ISO, officially began operations on February 23, 1947.

The ISO is the world's largest developer of technical standards. It is comprised of 148 countries, with one member per country, and has a Central Secretariat in Geneva, Switzerland, to coordinate its activities. Because any acronym for the organization would be different in the different languages of its member countries, the organization uses the name ISO, derived from the Greek *isos*, meaning "equal," for both its organization and the standards it issues. By providing the framework for compatible technology worldwide, ISO has helped build a world economy, making it easier to do business across national borders and providing customers with the benefits of a more competitive marketplace. ISO standards not only ensure compatibility but also address reliability and safety.³

2.2 ISO 9001: Quality Management System Requirements

As the European Trading Community began to take shape in the 1980s, there was a perceived need for a common quality standard for all nations. ISO assigned this task to Technical Committee 176, and in 1987, the ISO 9000 Quality System Standards were issued. Several revisions have been released since then, the most significant occurring with the 2000 release. This changed the standard from being a series of discrete elements to a systems-based approach—processes with inputs and outputs interacting with other processes. Key to this process approach is top management's commitment to plan and oversee quality throughout the organization. At the same time, the emphasis was increased on customer satisfaction resulting from high-quality, reliable products.

This more closely approximates how organizations actually operate, especially compared with the separate elements of the previous editions. The new standard was built around eight quality management principles, with increased emphasis on customer focus through requirements for measuring and tracking customer satisfaction, thereby focusing more on product quality than previous revisions. At this time, a new major revision is in preparation, revisiting the continued applicability of the quality management principles as well as possibly adding some requirements.

Because of the wide variety of operations intended to be covered by the ISO 9001 standard, not every requirement applies to every company. Therefore, Section 7 of the standard, covering product realization, can be tailored (within limits and with the concurrence of the registrar) to match the individual organization's operations.⁴

There are many similarities between ISO 9001's Quality Management Principles and the Core Values adopted from the Baldrige National Quality Award criteria to recognize performance excellence in national and local awards around the world. In giving guidelines for performance improvement and by addressing maturity models, ISO 9004 provides a link to Baldrige National Quality Award criteria and the CMMI maturity model (Table 1).

The primary publications in the ISO 9000 family include:

ISO 9000: Quality management systems—Fundamentals and vocabulary

ISO 9001: Quality management systems—Requirements

ISO 9004: Quality management systems—Managing for the sustained success of an organization—A quality management approach

ISO 19011: Guidelines for quality and/or environmental management systems auditing

2.3 ISO 9001: Certification/Registration

Separate from the ISO 9001 Quality Management System standard per se is the certification/registration process that has been institutionalized in many countries. The process requires a review by a third-party registrar of a company's documented quality system and the implementation of that system through on-site audits. The third-party registrar certifies that the system meets all of the requirements of the ISO 9001 model. The registration of the quality system can then be publicized. The registrar also performs regular reviews and periodic recertification audits.

The ANSI-ASQ National Accreditation Board (ANAB) is the U.S. agency that accredits agencies and individuals to serve as registrars. It assumed the accreditation duties of the Registrar Accreditation Board (RAB) when the RAB ceased operations in 2004. The ANAB covers the accreditation of Quality Management System (QMS) and Environmental Management System (EMS) registrars. Separate from the ANAB are certification programs for EMS auditors and QMS auditors and accreditation programs for course providers offering QMS and EMS auditor training courses.

The effort to obtain ISO 9001 registration typically takes 12–18 months from the time an organization makes the commitment to become registered until its quality system receives the certificate from its third-party registrar. The cost of registration varies depending on the size and complexity of the organization, the number of locations to be included on the registration certificate, and the state of its existing quality system when the decision to obtain registration is made.

Third-party registrars are generally contracted for three years. In addition to the initial assessment for registration, the registrar may be asked to perform a preassessment audit or to conduct training. Many companies find it helpful to hire an outside consultant to help prepare for ISO registration. There are many texts available on the subject of ISO 9001 quality systems and the registration process.⁵ (See Table 2.)

The "turtle diagram" is frequently used for mapping processes (Figure 1) in preparation for ISO 9001 certification. The process inputs and outputs become the turtle's head and tail, while the four corner boxes represent the legs and answer the questions: With What? With Whom? How Many? and How?

2.4 AS 9100: Quality Management Systems—Requirements for Aviation, Space and Defense Organizations

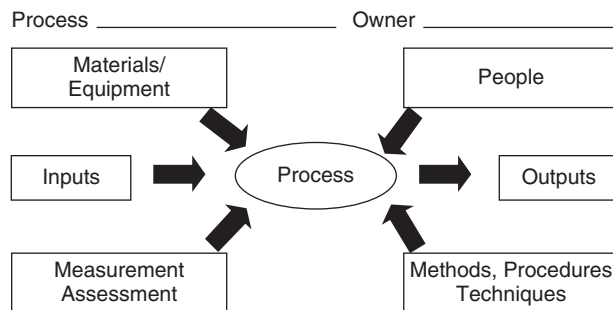
Since ISO 9001 was created as a quality management system standard applicable to nearly any organization, small or large, simple or complex, some industries felt that additional definition

Table 1 Comparison of ISO 9000 Quality Management Principles and Baldrige National Quality Award Core Values and Concepts.

ISO 9000 Quality Management Principles	Baldrige National Quality Award Core Values and Concepts
<p>Principle 1: Customer focus. Organizations depend on their customers and therefore should understand current and future customer needs, should meet customer requirements, and should strive to exceed customer expectations.</p>	<p>Customer-Driven Excellence. Customer-driven excellence is a strategic concept. It is directed toward customer retention, market share gain, and growth. It demands constant sensitivity to changing and emerging customer and market requirements.</p>
<p>Principle 2: Leadership. Leaders establish unity of purpose and direction of the organization. They should create and maintain the internal environment in which people can become fully involved in achieving the organization's objectives.</p>	<p>Visionary Leadership. An organization's leaders should set directions and create a customer focus, clear and visible values, and high expectations. Senior leaders should serve as role models through their ethical behavior and personal involvement.</p>
<p>Principle 3: Involvement of people. People at all levels are the essence of an organization and their full involvement enables their abilities to be used for the organization's benefit.</p>	<p>Visionary Leadership. Leaders should inspire and motivate the entire workforce and encourage all employees to contribute, to develop and learn, to be innovative, and to be creative.</p>
<p>Principle 4: Process approach. A desired result is achieved more efficiently when activities and related resources are managed as a process.</p>	<p>A process approach is expected but not expressly called out in the Core Values. Category 6 of the criteria, Process Management, considers key product/service and support processes.</p>
<p>Principle 5: System approach to management. Identifying, understanding, and managing interrelated processes as a system contributes to the organization's effectiveness and efficiency in achieving its objectives.</p>	<p>Systems Perspective. Successful management of overall performance requires organization-specific synthesis and alignment. A systems perspective means managing your whole organization, as well as its components, to achieve success.</p>
<p>Principle 6: Continual improvement. Continual improvement of the organization's overall performance should be a permanent objective of the organization.</p>	<p>Managing for Innovation. Innovation means making meaningful change to improve an organization's products, services, and processes and to create new value for the organization's stakeholders.</p>
<p>Principle 7: Factual approach to decision making. Effective decisions are based on the analysis of data and information.</p>	<p>Management by Fact. Organizations depend on measurement and analysis of performance. Such measurements should derive from business needs and strategy.</p>
<p>Principle 8: Mutually beneficial supplier relationships. An organization and its suppliers are interdependent and a mutually beneficial relationship enhances the ability of both to create value.</p>	<p>Valuing Employees and Partners. External partnerships might be with customers, suppliers, or education organizations. Strategic partnerships and alliances are increasingly important.</p>
<p>Not included in ISO 9000 Quality Management Principles</p>	<p>Agility, Public Responsibility and Citizenship, Focus on the Future, Focus on Results and Creating Value, Organizational and Personal Learning.</p>

Table 2 Number of Companies Registered To Various ISO Standards As of 2010

	ISO 9001	ISO 14001	ISO/TS 16949	ISO 22000	ISO 13485
Africa/West Asia	633,357	8,557	4,196	2,597	1,060
Central/South America	40,665	6,423	1,531	414	219
North America	36,632	6,302	5,217	181	4,040
Europe	530,722	103,126	10,624	7,083	11,034
Far East	428,755	124,922	22,215	8,263	2,401
Australia / New Zealand	9,784	1,642	163	92	80
Total	1,109,915	250,972	43,946	18,630	18,834

**Figure 1** Process maps are commonly used by certification bodies for visualizing a process and its links.

was required to ensure effective quality management for them. Like other similar standards, AS9100 includes the content of ISO 9001 in its entirety, with additional requirements inserted throughout. The standard is overseen by the International Aerospace Quality Group (IAQG). The most notable additions to AS 9100 not found in ISO 9001 include more detailed requirements for design and development functions and a documented configuration control system. (See Figure 2.) Product key characteristics are required to be identified, and the standard goes into detail on the topics of validation and verification, requiring verification documentation as well as validation test results.⁶

There are additional standards to help implement AS9100:

- AS9102: Aerospace First Article Inspection Requirement
- AS9103: Variation Management of Key Characteristics
- AS9110: Quality Management Systems—Requirements for Aviation Maintenance Organizations
- AS9120: Quality Management Systems—Requirements for Aviation, Space and Defense Distributors
- AS 9131: Quality Systems Non-Conformance Documentation

2.5 ISO/TS 16949: Quality Management Systems—Particular Requirements for the Application of ISO 9001:2008 for Automotive Production and Relevant Service Part Organizations

The automobile industry in the United States was the first to require its own version of ISO 9001 with industry-specific requirements, including a section of requirements for each of the Big

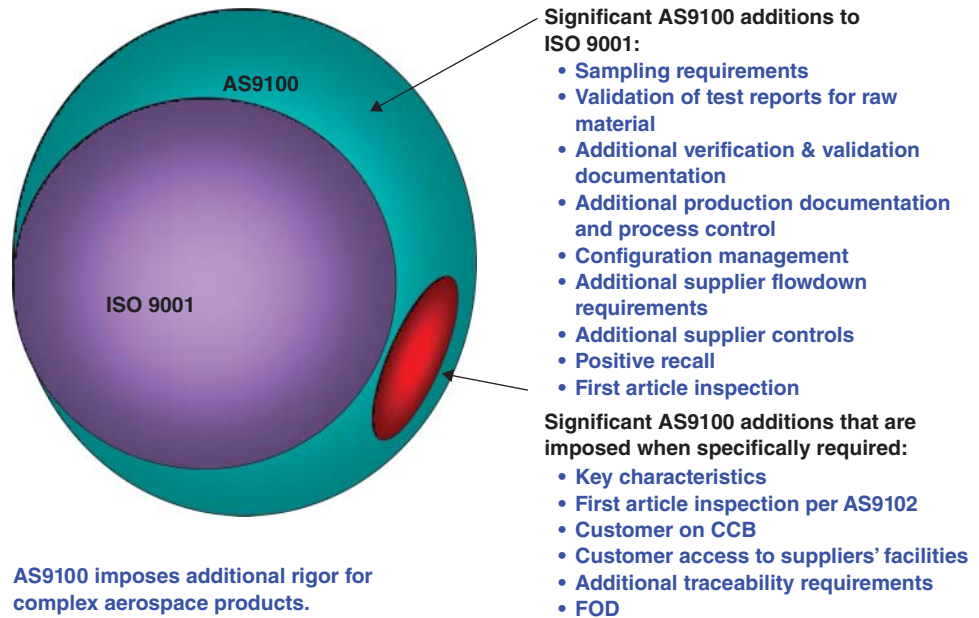


Figure 2 A comparison of the requirements shows how the entire ISO 9001 standard is not included in AS9100.

Three U.S. auto makers at the time (General Motors, Ford, and Daimler-Chrysler). QS 9000 was issued in 1994 by the Automotive Industry Advisory Group (AIAG). An international standard, ISO/TS 16949, was created by the International Automotive Task Force in collaboration with ISO in 1998, and revisions have been published to remain in alignment with the latest ISO 9001 revision. The standard applies to both original equipment manufacturers (OEMs) and suppliers of automotive parts.

ISO 9001 is the central tenet of ISO/TS 16949; however, the automotive standard includes additional requirements. Since the standard applies throughout the supply chain, it offers enhanced commonality of processes, which in turn allows for greater confidence in global sourcing and elimination of multiple sources. In fact, for a company to be considered as a supplier to the international automotive market, it must be listed on customers' potential supplier list and achieve registration to ISO/TS 16949. Like ISO 9001, ISO/TS 16949 requires a process approach (see Table 1) and places greater emphasis on management commitment, customer satisfaction, measurement, and internal auditor qualifications.⁷

Oversight and certification is managed by the International Automotive Task Force, www.iaatfglobaloversight.org. To compound suppliers' level of difficulty in complying with ISO/TS 16949, major automotive manufacturers have issued addendums with company-specific requirements.⁸

2.6 ISO 13485: Medical Devices: Quality Management Systems—Requirements for Regulatory Purposes

In order to ensure a common quality system for organizations that design, develop, produce, install, and service medical devices, the ISO 13485 Medical Devices standard was issued in 1996 and has been revised and updated subsequently to align with ISO 9001. Notable additions to the ISO 13485 standard include awareness of regulatory requirements as a management

responsibility, additional focus on risk management and product safety, and additional rigor required in the areas of process validation, traceability, and effectiveness of corrective and preventive action.

Registration to the requirements of ISO 13485 is increasingly becoming a requirement of doing business in the medical devices industry. An audit of the finished device manufacturer will examine the capabilities of the entire supply chain, including ISO 13485 registrations throughout the chain. Registration to ISO 13485 does not fulfill the requirements of industry regulators but is generally the foundation of those regulatory requirements.⁹

2.7 TL 9000: Telecom Quality Management System

TL 9000 is the telecommunication industry's equivalent of ISO 9001 quality management system requirements. It was developed by the Quality Excellence for Suppliers of Telecommunications (QuEST) Forum. Like the other industry-specific standards, TL9000 uses ISO 9001 as its foundation, adding requirements unique to the telecommunications industry. TL 9000 includes over 90 enhancements dealing specifically with telecommunications-related issues, driving improvements in that industry. One major difference between TL 9000 and other industry-specific quality management system standards is that common, industry-related metrics are required. In addition to the required internal metrics, TL9000 requires quarterly submission of the metrics to a central repository at the University of Texas at Dallas. This requirement allows for industry benchmarking.

TL 9000 is comprised of two primary documents:

- TL 9000 Requirements Handbook—the ISO 9001 standard with industry-specific requirements added
- TL 9000 Measurements Handbook—the document defining the metrics to be collected and submitted for industrywide use and comparison

The benefits of utilizing a TL 9000 quality management system are becoming more widely recognized, along with the value of having industrywide metrics. The telecommunications industry was initially slow to accept TL 9000 certifications, but international pressures are continually driving companies to use TL 9000 to ensure their competitiveness in the market.¹⁰

2.8 Mission Assurance—Missile Defense Agency Assurance Provisions

The term *mission assurance* has been embraced by the U.S. National Aeronautics and Space Administration (NASA) and the U.S. Missile Defense Agency (MDA) to encompass the development, engineering, testing, production, procurement, and implementation of space vehicles or missile defense elements under the cognizance of NASA or the MDA.

In its quest for mission assurance, NASA developed the Process Based Mission Assurance (PBMA) plan in 2002. In early 2004, the MDA issued the mission assurance provisions (MAPs) to provide a measurable, standardized set of safety, quality, and mission assurance requirements to be applied to contracts for mission and safety-critical items in support of evolutionary acquisition and deployment of MDA systems.

Both programs are predicated on the idea that if characteristics critical to mission success are identified in designs/hardware/software and measures are taken to ensure that these critical characteristics are met, the end product will function successfully. Mission assurance requires zero defects in those items critical to the safety, reliability, and quality of the end product.¹¹

Unlike the standards discussed above, the MAP is not directly based on ISO 9001. While the MAP and ISO 9001 (and the various standards derived from ISO 9001) are generally compatible, the MAP includes significantly more detail in its requirements. For instance, the one area which deviates the most involves safety. AS9100 discusses safety only obliquely relating

to product safety—control of foreign objects that may cause damage, for instance. The MAP, however, interprets safety as “system safety” and offers extensive guidance on expectations and requirements to ensure success of the overall mission.

A third-party audit and certification program has not been developed for mission assurance. However, MAP requirements are being included in certain MDA contracts, resulting in these requirements being flowed down to subtier suppliers.

2.9 Capability Maturity Model Integration

The Capability Maturity Model Integration (CMMI®) is an outgrowth of an earlier software engineering capability maturity model (SE-CMM) contracted by the DoD in collaboration with Carnegie Mellon University (CMU) in Pittsburgh, Pennsylvania, in the early 1980s. The Software Engineering Institute (SEI) is a federally funded research and development center sponsored by the DoD and operated by CMU. The SEI was established to address growing software costs and quality problems. Maturity levels for software development were documented, and audits were conducted to verify that government suppliers had reached specific levels of maturity¹² (Fig. 3).

Since then, CMMI® has grown to include three general areas of interest:

1. Product and service development—CMMI-DEV
2. Service establishment, management and delivery—CMMI-SVC
3. Product and service acquisition—CMMI-ACQ

Each of these assesses an organization’s maturity in a group of processes. CMMI® offers two representations: continuous and staged. In the continuous model, the organization focuses on specific processes deemed to be important to the business. The staged representation offers a sequence of improvements that utilize each of the processes. See Fig. 3 for a representation of the maturity levels assessed via CMMI®. The DoD has begun requiring at least CMMI® level 3 for some of its contracts, and requirements for levels 4 and 5 compliance are becoming more common. An organization must undergo a Standard CMMI® Appraisal Method for Process Improvement (SCAMPI) in order to have its maturity level recognized.

2.10 ISO 14000: Environmental Management System Requirements/EMAS

The ISO 14000 series of environmental management standards was released in 1996 and most recently revised in late 2004, with a Technical Corrigendum in 2009. The standards represent the work of the ISO’s Technical Committee 207.

The ISO 14000 family of standards is primarily concerned with environmental management, focusing on how an organization minimizes harmful effects of its products, processes, and services on the environment and how that organization achieves continual improvement of its environmental performance.¹³

The ISO 14001 registration process is similar to that of the quality management system standard, ISO 9001. The ANAB serves as the U.S accrediting body, and many of the same registrars that audit and certify organizations to ISO 9001 are accredited to serve as registrars for ISO 14001.

In fact, ISO 19011 (Guidelines for Quality and/or Environmental Management Systems Auditing) allows for a common auditing process for both ISO 9001 and ISO 14001 systems.

General requirements for ISO 14001 include:

- Establishment of an environmental management system, including an environmental policy, objectives and targets, and identification of the environmental aspects of the organization’s activities, products, and services.

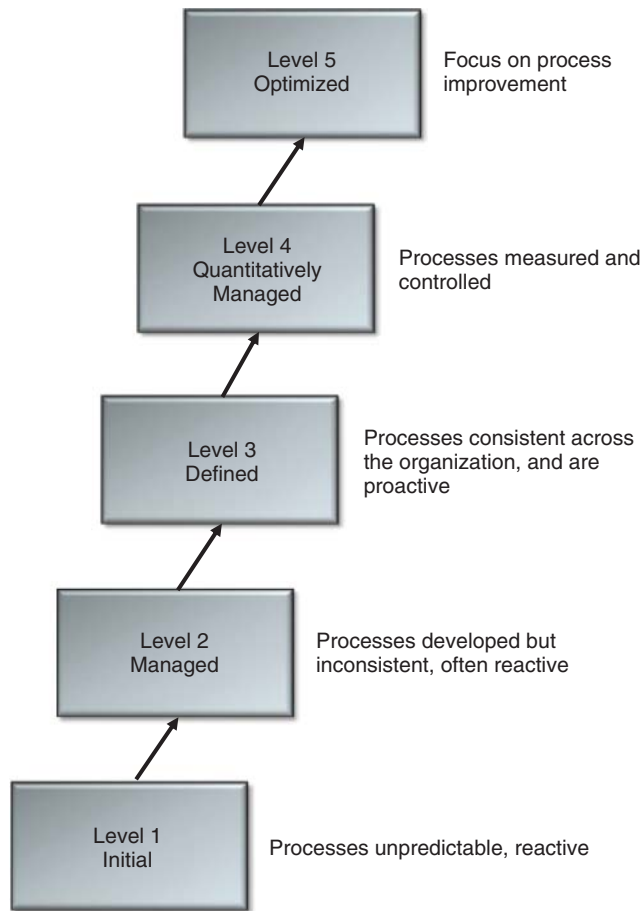


Figure 3 Representation of how an organization's approach to processes varies as maturity levels increase. It is of course possible for different processes to operate at different levels of maturity

- Documented procedures as well as documents and records as required by the EMS standard and regulatory agencies.
- Evidence that procedures are being followed and the EMS maintained. This may include internal audits and periodic management reviews.
- Corrective and preventive action processes and records.
- Continuous improvement processes.¹⁴

Although there was speculation when ISO 14001 was first issued in 1996 that findings in the registration process might result in sanctions from the Environmental Protection Agency (EPA) or, conversely, that registration to ISO 14001 might lead to decreased oversight by the EPA, neither event seems to have occurred.

Rather than focusing on specific compliance with environmental regulations (which are often dictated by governmental bodies), ISO 14001 offers an outline of what an organization must do to effectively manage and control its environmental management system. Like the

other standards mentioned here, ISO 14001 is arranged around the Shewhart cycle—plan, do, check, and act:

- Plan your processes to establish objectives and processes.
- Do what is necessary to implement those processes.
- Check the compliance and effectiveness of those processes.
- Act to improve the processes based on those results.

The European Union's environmental management scheme EMAS builds on ISO 14001 requirements, adding measurement and evaluation requirements, legislative compliance, and employee involvement requirements in particular. EMAS registration has quickly become globally applicable, with thousands of sites now EMAS registered.¹⁵

2.11 ISO 22000: Food Safety Management Systems—Requirements for Any Organization in the Food Chain/HACCP

The principles of hazard analysis and critical control points (HACCP) offers a systemic approach to food and pharmaceutical safety. Utilizing preventive action methodologies, it affords ways to ensure the prevention of hazards in food rather than inspecting the finished product.

HACCP principles grew out of an effort by NASA in the 1960s to ensure food safety for space flight. Its approach to preventing problems rather than removing them after the fact has begun to spread to other industries besides food, including pharmaceuticals.

HACCP is a management system in which food safety is addressed through the analysis and control of biological, chemical, and physical hazards from raw material production, procurement, and handling to manufacturing, distribution, and consumption of the finished product. For successful implementation of a HACCP plan, management must be strongly committed to the HACCP concept. A firm commitment to HACCP by top management provides company employees with a sense of the importance of producing safe food.

HACCP is designed for use in all segments of the food industry from growing, harvesting, processing, manufacturing, distributing, and merchandising to preparing food for consumption. Prerequisite programs such as current Good Manufacturing Practices (cGMPs) are an essential foundation for the development and implementation of successful HACCP plans. Food safety systems based on the HACCP principles have been successfully applied in food processing

Table 3 Comparison of Management System Standards

Standard	Industry	Includes ISO 9001 Requirements?	U.S. Oversight or Accreditation Body
ISO 9001	Generic quality management system	N/A	www.anab.org
ISO/TS 16949	Automotive	Yes	iatfglobaloversight.org
AS 9100	Aerospace	Yes	www.anab.org
ISO 13485	Medical devices	Yes	www.anab.org
TL 9000	Telecommunications	Yes	www.anab.org
Mission assurance	Missile defense	No	Missile Defense Agency, no accreditation
CMMI	Generic maturity model	No	www.sei.cmu.edu/cmmi/
ISO 14001	Generic environmental management system	No	www.anab.org
ISO 22000	Food safety management	No	www.anab.org

plants, retail food stores, and food service operations. The seven principles of HACCP have been universally accepted by government agencies, trade associations, and the food industry around the world.¹⁶

In 2005, the ISO published a standard documenting the HACCP model—ISO 22000. Similar to the other ISO standards, a system of registration/certification developed to recognize companies' efforts to meet the requirements of the standard. In the United States, HACCP programs have become mandatory for meat, seafood, and juice and are currently voluntary in other industries.

Table 3 provides a comparison of the more frequently used management system standards.

3 QUALITY AND PERFORMANCE EXCELLENCE AWARDS

Awards for quality and performance excellence have played an important role in the worldwide improvement of organizational performance over the past 30 or more years. Although an award is presented in recognition of an organization's accomplishments, it is often the award criteria that drive world-class performance. For most organizations, the cycle time averages three years from beginning an effort to pursue a performance excellence award such as the U.S. Baldrige National Quality Award, European Quality Award, or other Baldrige-based national and state awards to achieving that award. This cycle includes annual award applications and improvement activities based on feedback. In the case of the Deming Prize, this preparation period may take five years or longer.

Award criteria are updated regularly to reflect the best practices of successful organizations as well as emerging challenges. For example, after a number of business scandals in the United States in the late 1990s, corporate responsibility and citizenship received greater emphasis in the Baldrige Award criteria. The requirement for data gathering and analysis in early award criteria has been expanded to include knowledge management.

Organizations pursue awards for a variety of reasons. Many awards place heavy emphasis on measurable results, positive trends, and world-class performance in comparison with competitors and best-in-class organizations. For businesses, this emphasis drives increased revenue and market share. In health care, results include better patient outcomes, and in education, higher test scores. In most cases, operating costs are reduced. Customer satisfaction is key to any organization's success, and it takes a prominent place among results categories. Whether an organization pursues an award for its prestige and positive publicity or to improve its processes, the pursuit can be rewarding whether the award itself is ever attained.

3.1 Deming Prize

The Deming Prize was established in 1950 by the Japanese Union of Scientists and Engineers (JUSE). It was named after U.S. statistician Dr. W. Edwards Deming to recognize his contributions to Japanese quality control. Deming was invited to Japan in 1950 to present a series of lectures on quality control and statistical techniques. At the time Japan was still occupied by Allied forces and the Japanese were beginning to rebuild their industries. Deming's approach to quality control was instituted throughout Japan. It was later broadened to include TQM, although Deming disavowed any relationship to TQM. Deming's lectures and the prize named after him launched a movement that transformed Japan's standing in world markets, earning the reputation as premier quality leader, particularly in electronics and automobiles.

There are four types of Deming awards (see Table 4). The Deming Prize is open to companies or autonomous divisions that have achieved distinctive performance improvement through the application of TQM in a designated year. Applicants must complete a formal application in Japanese. The JUSE Deming Prize Committee evaluates each application and, if there is enough evidence to move forward, completes a document review. A site visit is also required, and costs for travel from Japan are paid by the applicant.¹⁷

Table 4 Categories of Deming Prize

Type of Prize	Candidate	Description
The Deming Grand Prize (formerly the Deming Metal)	Organizations may apply three or more years after receiving the Deming Prize and may continue to apply every three years	Recognizes organizations where TQM has improved substantially beyond the level at the time they won the Deming Prize
The Deming Prize	Organizations or divisions of organizations that manage their business autonomously	Given to organizations or divisions of organizations that have achieved distinctive performance improvement through the application of TQM in a designated year
The Quality Control Award for Operations Business Units	Operations business units of an organization	Given to operations business units of an organization that have achieved distinctive performance improvement through the application of quality control/management in the pursuit of TQM in a designated year
The Nikkei QC Literature Prize	Author(s) of literature published in Japan	Literature on the study of TQM or statistical methods used for TQM that is recognized to contribute to the progress and development of quality management

The JUSE Committee uses a checklist that includes:

- Policies
- Organization
- Education
- Collection
- Analysis
- Standardization
- Control
- Quality assurance
- Effects (results)
- Future plans

It is interesting to note that the Deming Prize Committee evaluates not only the results that have been achieved by the applicant but also the effectiveness expected in the future. The criteria do not specify how the applicant should approach TQM, and the committee evaluates how the applicant uses TQM in the context of its unique business situation. In its publication concerning the evaluation process, the committee states¹⁸:

No organization can expect to build excellent quality and management systems just by solving problems given by others. They need to think on their own, set lofty goals, and challenge themselves to achieve these goals.

The Deming Prize involves a process that can take several years and cost a great deal. Implied in this process is the use of JUSE consultants for months or years to assist the applicant in putting its TQM processes in place. The consultants perform a quality control diagnosis and recommend changes.

The organization creates its application for the Deming Prize the year after the JUSE consultants have completed their work. The length of the application is set according to the size of the company, ranging from 50 pages for organization with fewer than 100 employees to 75 pages for 100–2000 employees, plus 5 pages for each additional 500 employees over 2000.

The Deming Prize has been awarded to over 200 companies since 1951. Although the competition is open to any company regardless of national origin, most of the recipients in the early years of the award were Japanese, and many from outside the company continue to have parent companies in Japan. Only three U.S. companies have received the Deming Prize. Florida Power and Light was the first U.S. company to receive the prize in 1989, AT&T Power Systems was a recipient in 1994, and Sanden International, Inc. received the award in 2006. Since 2001 companies in Thailand and India have dominated the recipient list.

3.2 Baldrige National Quality Award

Although not the oldest quality award, the Baldrige National Quality Award (BNQA) program has had the greatest influence on performance excellence in the United States. It was created by the U.S. Congress as Public Law 100-107 on August 20, 1987, to address international trade deficits and a lagging U.S. economy caused in large part by the availability of Japanese automobiles and electronics perceived to be of higher quality and reliability than their U.S. counterparts. The Baldrige Award was named after Malcolm Baldrige, the U.S. Secretary of Commerce who died in a tragic rodeo accident in 1987.¹⁹

The Baldrige Award program has four roles in strengthening U.S. competitiveness:

1. Raises awareness about the importance of performance excellence in driving the U.S. and global economy
2. Provides organizational assessment tools and criteria
3. Educates leaders in businesses, schools, health care organizations, and government and nonprofit agencies about the practices of best-in-class organizations
4. Recognizes national role models and honors them with the only Presidential Award for performance excellence

The award was first offered only to U.S. for-profit companies in the categories of manufacturing, service, and small business (fewer than 500 employees) with a maximum of two award recipients per year in each category. The award was later amended to increase the maximum number of recipients to three per category, and in 2000, education and health care were added. In 2004, Congress passed legislation to add not-for-profit and government organizations to those eligible to apply for the Baldrige Award.

The Department of Commerce is responsible for administering the BNQA program. The National Institute of Standards and Technology (NIST), an agency of the Department of Commerce, manages the award program. The American Society for Quality (ASQ) assists in administering the program under contract to NIST.

There are seven criteria *categories*. Each is broken down into subcategories called *items*, and each Item has a number of *areas to address* that are comprised of specific questions about the applicant's processes or results. An application is scored on a 1000-point weighted scale, with points assigned to each Item. Figure 4 provides a breakdown of points by category. The criteria, including weighting of the categories and points per item, are revised periodically. At this writing, separate criteria are used for business/nonprofit, health care, and education organizations; however, the differences between them are small, and there is a plan to eventually use common criteria for all applicant groups.²⁰

The results-oriented BNQA criteria is built around 11 core values and concepts (see Fig. 5) and focuses on an organization's primary activities, customers, and competitive results. The

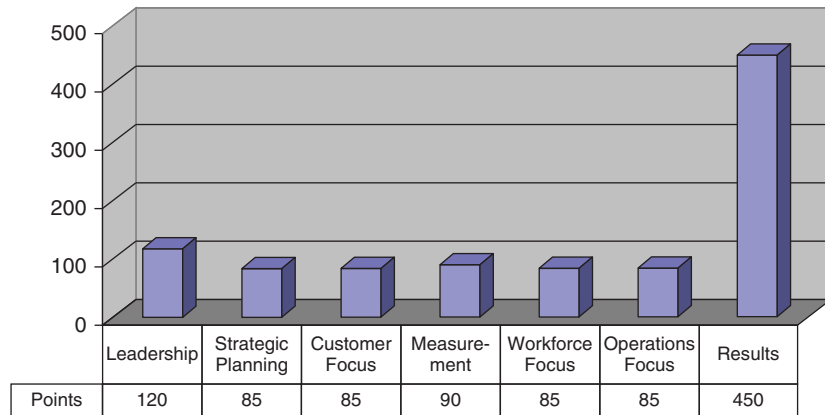


Figure 4 Baldrige criteria and point values.

The Role of Core Values and Concepts

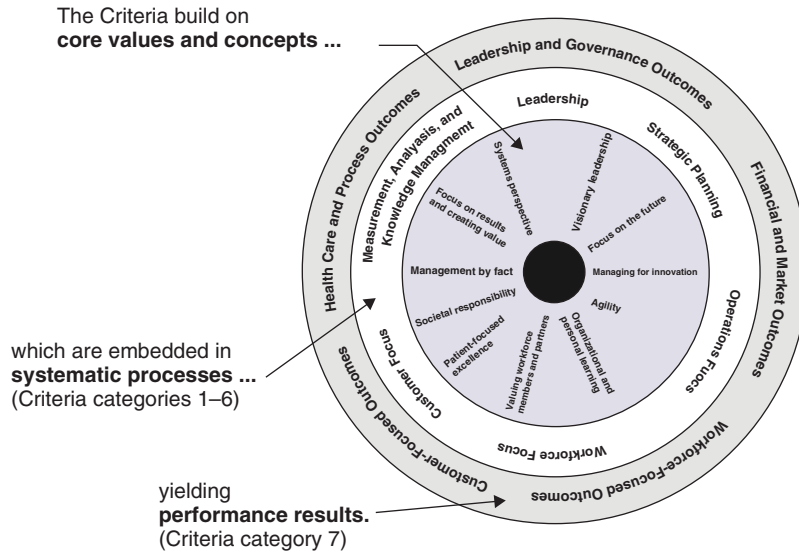


Figure 5 Core values and concepts support performance and results in the BNQA criteria.

greatest changes to the criteria were made in 1995, when the word quality was almost entirely removed, broadening the scope of the award criteria to encompass all of the elements of an organization’s performance and not just TQM. Quality management systems must be fully integrated into an organization’s operations.

The Baldrige scoring system is based on two evaluation dimensions: (1) process and (2) results. “Process” refers to the methods the organization uses and improves to address the item requirements in categories 1–6. The four factors are used to evaluate the process are approach, deployment, learning, and integration (ADLI).

The ADLI model promotes an organizational focus and places emphasis on processes that help the organization share its knowledge and information. Improvement outcomes are

expected to be integrated across the organization’s operations, thereby improving overall organizational performance.

In addition to “process,” the Baldrige scoring system has a separate set of scoring factors for “results.” Results are defined as the organization’s outputs and outcomes in achieving the requirements called out in criteria items 7.1–7.5. The four factors used to evaluate results are levels, trends, comparisons, and integration (LeTCI). Results items call for data showing performance levels, trends, and relevant comparisons for key measures and indicators of organizational performance and integration with key organizational requirements. For example, if the organization states that a key requirement is to grow the business by 20% each year, the results category should include a chart showing the current growth level, the trend from past years, and a comparison with growth levels for best-in-class organizations, competitors, or other relevant organizations.²¹

Figure 6 illustrates the BNQA framework. The Preface: Organizational Profile sets the context in which the organization operates. The system operations are composed of the six Baldrige categories in the center of the figure. These define the organization’s operations and the results they achieve. Leadership (category 1), strategic planning (category 2), and customer focus (category 3) represent the leadership triad. Workforce focus (category 5), operations focus (category 6), and results (category 7) represent the results triad in that employees and key processes accomplish the work of the organization that yields the business results.

Measurement, analysis, and knowledge management (category 4) serves as the system foundation in the Baldrige model. These are critical to the effective management of the organization, enabling a fact-based, knowledge-driven system for improving performance and competitiveness.²⁰

A board of examiners is selected each year to review applications for the BNQA. Board members are a volunteer group of recognized experts in the areas of performance excellence

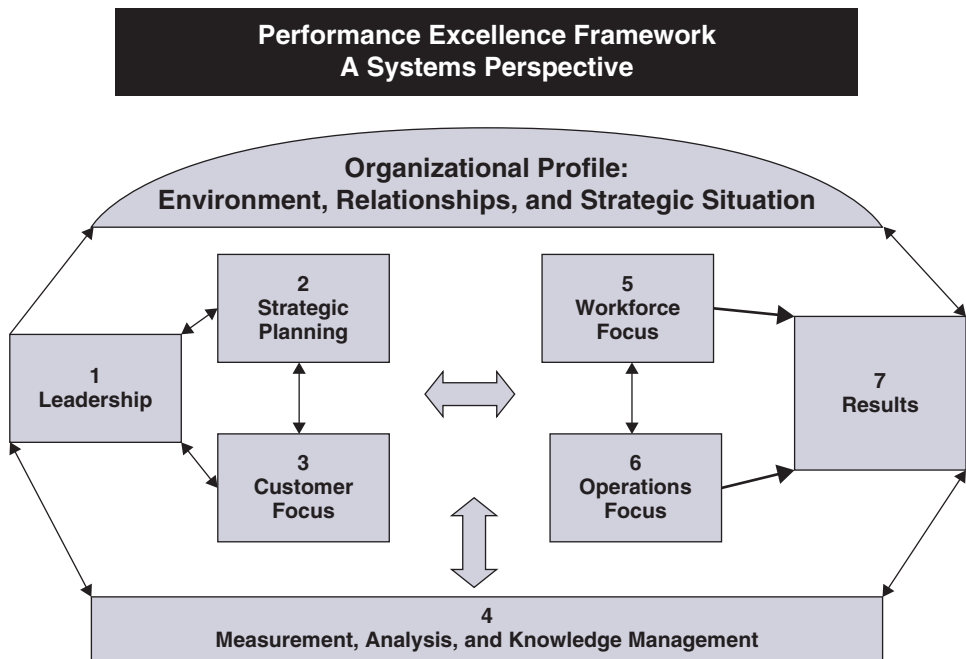


Figure 6 BNQA criteria framework with breakdown of points per category.

and continuous improvement. They are selected through an application process, with at least one-third of the board made up of new examiners. An examiner may serve for no more than six years and then takes on alumni status. Alumni examiners may apply for a seat at examiner training and can be called upon to review an application during the first stage in the award process.

The board has almost doubled in size since the award's inception to accommodate a spike in applications after education, health care, and nonprofit/government organizations were added to the award categories. In 2012, there were over 1000 volunteers working within the Baldrige program, many of whom were examiners. Examiner applications are due in early January each year and are available at www.baldrige.nist.gov each Fall.

Applications for the BNQA are limited to a five-page organizational profile and a 50-page response to the criteria's areas to address. The award process includes:

- Stage 1: Seven to 10 examiners independently evaluate the application.
- Stage 2: Examiner team from stage 1 work together to create a consensus scorebook; judges review consensus scores and select applicants to receive a site visit; a subset of the team prepares the feedback report for each applicant that does not move on to stage 3.
- Stage 3: Examiner teams visit sites for two to five days, depending on the applicant's size and complexity. The team prepares a final scorebook with recommendations for changes in the original score; judges select applicants to receive the award; a subset of the examiner team prepares the feedback report for each applicant that has received a site visit.

Every applicant receives a detailed feedback report citing strengths and opportunities for improvement on each of the criteria items. The report also contains an executive summary and an explanation of the application's scoring band and what that may indicate about the organization in terms of the ADLI model.

Applications are due at the end of May each year. Stage 1 activities take place in June to July and stage 2 in August to September. Site visits take place in October, and award recipients are announced in December.

The application fees for the BNQA range from \$1750 for not-for-profit educational organizations to \$7500 for large manufacturing and service businesses. Complex organizations with more than one product line may opt (with prior permission) to provide supplemental information. The fees for submitting this additional material range from \$250 to \$2000. Site visit costs, reimbursed by the applicant, can range from as little as \$1500 to as much as \$40,000. These amounts will most likely increase over time.²²

The Baldrige program has developed from an annual award process to a more comprehensive support service for organizations that seek to improve their operations. Countless studies have demonstrated the value of the BNQA to U.S. organizations. In a report published by the U.S. Department of Commerce, 273 BNQA applicants between 2006 and 2011 were surveyed. The benefit-to-cost ratio was found to be 820:1.²³ In a Thomas Reuters study of the 100 top hospitals, Baldrige hospitals outperformed non-Baldrige hospitals on all of the measures used, including patient mortality, safety, and length of stay.²⁴

Unfortunately, economic pressures in the United States along with some political factions produced a situation in 2012 where the entire budget for the Baldrige program was cut. Despite such a drastic blow, there were enough citizens and organizations that believed in the value of the program that funding was raised for it to continue. At this time, the Baldrige program is a public-private program dedicated to performance excellence.²⁵ The Baldrige organization has performed a major overhaul to the program, reducing staff and budgets while changing the eligibility requirements and leveraging state and local award programs. An applicant for the

BNQA must have either received the Baldrige Award in the past or have received a state or local award.

3.3 U.S. State Quality Awards

Since the introduction of the BNQA, virtually all U.S. states and most of its territories have initiated state awards for performance excellence. Most of these awards are based on the Baldrige criteria, and in most cases eligibility has been extended to any organization within the state, including governmental and not-for-profit organizations.

The Alliance for Performance Excellence is a nonprofit network of national, state, and local Baldrige-based award programs that works closely with the BNQA and the American Society for Quality. With the financial challenges facing the Baldrige program, The Alliance will become a powerful screening step for future Baldrige applicants. The BNQA page of the NIST website contains a link to the complete listing of these awards programs.²⁶

State award programs generally adhere to the BNQA process with a written application, site visits, and an award ceremony. Applicants receive feedback reports. State examiners are selected and receive training, often using BNQA training materials made available by NIST. Unlike the Baldrige Award, most state programs have varying levels of recognition. This serves to encourage organizations that may not yet have achieved the highest levels of performance excellence. The costs to state award applicants are often less than half the BNQA costs, particularly the costs for a site visit. Many state quality award recipients have gone on to become recipients of the Baldrige Award.

3.4 Shingo Prize for Operational Excellence

The Shingo Prize for Operational Excellence was created in 1988, and named for Japanese industrial engineer Shigeo Shingo, a leading expert on improving concepts, management systems, and improvement techniques that have become known as the Toyota Business System. Shingo created, along with Taichi Ohno, many of the facets of just-in-time manufacturing while working with Toyota Production Systems. He is known for his books, including *Zero Quality Control; Source Inspection and the Poka-Yoke System; and The Shingo System for Continuous Improvement*.

The award is administered by the College of Business, Utah State University, and a Board of Governors made up of leading representatives of businesses, professional organizations, and academic institutions. The Board oversees fund raising and other financial activities, guides prize governance, establishes prize guidelines, and ratifies prize recipients based upon the recommendations of the Board of Examiners.

The mission of the Shingo Prize is to create excellence in organizations through the application of timeless, universal, and self-evident principles of operational excellence; alignment of management systems; and the wise application of improvement techniques across the entire organizational enterprise.

The Shingo Prize is awarded to organizations that demonstrate a culture where principles of operational excellence are deeply embedded into the thinking and behavior of all leaders, managers, and associates. The award has changed dramatically since its inception, when it was called the Shingo Prize for Excellence in American Manufacturing and was open only to manufacturers in the United States, Canada, and Mexico. Today the Shingo Prize is now open to virtually any organization anywhere in the world. Prize recipients have come from the United States, Mexico, Brazil, India, Denmark, and many other countries.

Changes to the Shingo model have moved its focus from the deployment of quality tools and techniques to the integration of guiding principles that lead to organizational excellence.

These Shingo Principles of Operational Excellence are illustrated in a figure called The Shingo House.

The house is made up of four levels:

The Cultural Enablers

Continuous Process Improvement

Enterprise Alignment

Results

Performance is measured both in terms of business results and the degree to which business, management, and work systems are driving appropriate and optimum behavior at all levels. The principles of operational excellence must be deeply imbedded into the organization's culture and regularly assessed for improvement.²⁷

3.5 Quality Awards around the World

It is a testament to the effectiveness of award programs at increasing organizational effectiveness that over one hundred regional and national performance excellence award programs have been established. There are now award programs in Europe, Japan, the U.K., South Africa, Egypt, Hong Kong, India, Philippines, Australia, Singapore, and countless other regions and nations. The criteria for these programs are most often based on the BNQA criteria, and NIST in the United States has been generous in sharing its training program and providing support to these programs.

The European Excellence Award was created in 1991 with the first awards made in 1992. The award is managed by the EFQM, formerly known as the European Foundation for Quality Management (www.eqfm.org), an organization founded in 1988 and made up of more than 440 quality-oriented European businesses and organizations. It was created to enhance European competitiveness and effectiveness through the application of TQM principles in all aspects of organizations. The EFQM headquarter is located in the Netherlands.

The EFQM is dedicated to fostering organizational excellence, whether private, public, or nonprofit. Three types of recognition are announced each year. The EFQM Excellence Award is the top award, recognizing outstanding performance across the EFQM Excellence Model. To win the award, an applicant must be able to demonstrate not only that their performance exceeds that of their peers but also that they will maintain this advantage into the future.

Prizes are awarded to organizations that demonstrate role model behavior in one of the following eight criteria²⁴:

- Leading with vision, inspiration, and integrity
- Managing processes
- Succeeding through people
- Adding value for customers
- Nurturing creativity and innovation
- Building partnerships
- Taking responsibility for a sustainable future
- Achieving balanced results

The prize may be awarded to an applicant based on outstanding performance in one or more of the role model behaviors. In addition to the awards and the prizes, a finalist distinction

is given to any organization that submits an application and scores more than 400 points based on a 1000-point scale.

The European Quality Award criteria are weighted and scored on a scale of 0–1000 points in a manner similar to the criteria for the Baldrige Award. The criteria are divided into two main categories: enabler criteria and results criteria. The EQA model is currently undergoing a major revision expected to be deployed in 2006.²⁸

Although Japan’s Deming Prize had been in existence for almost 40 years when the Baldrige Award was created, it remained focused on specific techniques associated with TQM. When the Japanese witnessed the improved quality of U.S. products and operations in the early 1990s, the Japan Quality Award (JQA) program was launched. The JQA was established in 1995 by the Japan Productivity Center for Socio-Economic Development (JPC-SED). It is modeled after the BNQA, modified to accommodate Japanese management practices.²⁹

The models for the MBQA and JQA are similar; however, it is interesting to note some of the differences reflective of the priorities set by management systems in the two countries. While the Baldrige criteria has taken a closer look at social responsibility through areas to address in the leadership category, the Japanese model (Fig. 7) includes an entire category for social responsibility. This mirrors the difference in registration levels to ISO 14001 Environmental Management Standard between Japan and the United States.

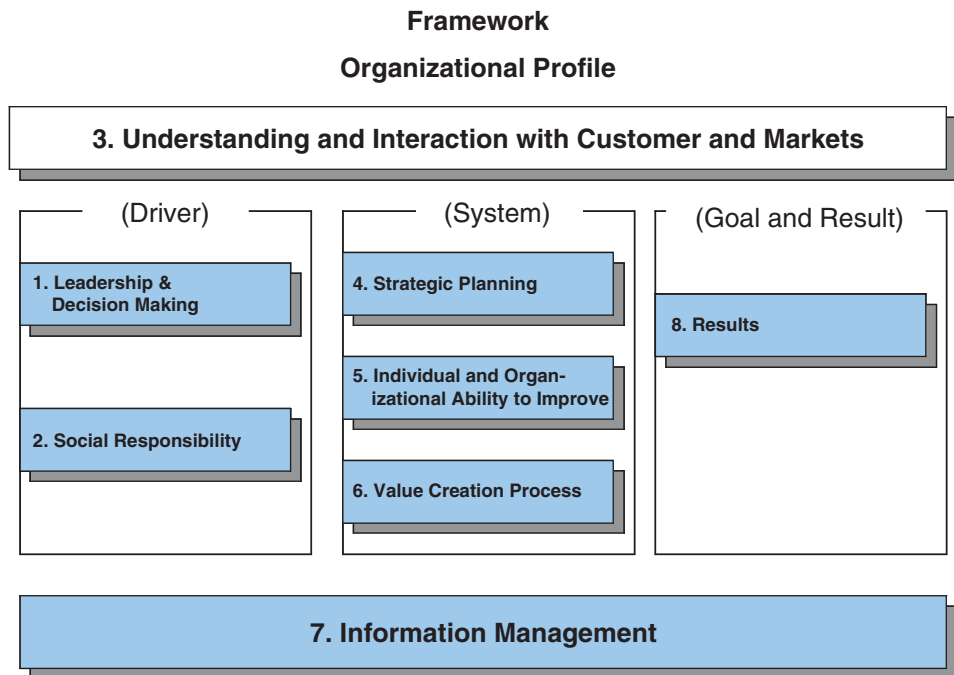


Figure 7 Criteria framework for the Japan Quality Award.

Table 5 Comparison of Quality Awards and Certifications, Considering Emphasis, Cost, and Weight of Social Responsibility Criteria

Award/ Certification	Year Created	Emphasis	Cost	Social Responsibility Requirements	Contact
Baldrige National Quality Award	1987	Performance excellence, customer focus, results	Medium to High	Medium	www.baldrige.nist.gov
Deming Prize	1950	TQM	Very High	Low	www.juse.or.jp/e/deming/
State Quality Awards (U.S.)	Various	Modeled on BNQA	Low-Medium	Medium	http://www.baldrigepe.org/alliance/
Shingo Prize	1988	Performance excellence, customer focus, results	Medium	Medium	www.Shingoprize.org
Japan Quality Award	1994	Performance excellence, customer focus, results	High	High	www.jpq-scd.or.jp/eng/award/
European Quality Award	1992	Performance excellence, customer focus, results	High	High	www.efqm.org
<i>Industry Week</i> Best Plants Award	1990	Manufacturing methods	Low	Low	www.industryweek.com
ISO 9001	2000	Quality management system	High	Low	www.iso.org
ISO 14001	2002	Environmental management system	High	High	www.iso.org

3.6 Industry-Specific Quality Awards

In addition to national and regional quality awards, there are a few significant award programs that address specific sectors or industries. The American Health Care Association Quality Award is given for excellence in long-term care facilities. The award is modeled after the BNQA.³⁰

The National Housing Quality (NHQ) Award gives the highest recognition by the housing industry for quality achievement. The NHQ awards are also patterned after the Baldrige National Quality Award. Entries are judged by panels of experts who evaluate the role that customer-focused quality plays in construction, business management, sales, design, and warranty service (National Association of Homebuilders, www.nahb.org).

Industry Week magazine's America's Best Plants Award was first presented in 1990. The award is given to 10 manufacturing plants in the United States, Canada, or Mexico that score the greatest number of points on a questionnaire. Independent judges select 25 semifinalists, and those applicants complete additional information. A site visit is performed by one of the magazine's editors at each of the 10 finalists' plants. Articles about the finalists appear in an issue of *IndustryWeek* magazine; however, there is no formal feedback to the applicants.³¹

3.7 How Do They Compare?

The value in pursuing registrations, certifications, or awards is not necessarily in achieving the certificate or plaque. The benefit is often derived from the process itself, particularly when an organization must apply for several years, receiving feedback from the awarding or certifying organization, and implementing improvements. While many of the awards focus on results, not all of them do. Table 5 provides contact information as well as some points of comparison between some of the award and certification programs.³²

REFERENCES

1. International Organization for Standardization, "Introduction and Overview," www.iso.org.
2. International Organization for Standardization, "Certification, Registration, and Accreditation," www.iso.org.
3. International Organization for Standardization, "Introduction," www.iso.org.
4. E. Stapp, *ISO 9000:2000, An Essential Guide to the New Standard*, Quality Publishing, Tucson, AZ, 2001.
5. D. Drickhamer, "Quality-Management Principles," *IndustryWeek*, March 5, 2001.
6. E. Barker, "Aerospace's AS9100 QMS Standard," *Quality Digest*, May 2002.
7. R. O'Connell, "ISO/TS 16949 Made Easy," *Quality Digest*, August 1, 2004.
8. C. Lupo, "ISO/TS 16949 the Clear Choice for Automotive Suppliers," *Quality Progress*, October 2002, p. 44.
9. J. Adam, "ISO 13485 Levels the Playing Field," *Quality Digest*, August 15, 2004.
10. R. Clancy, "Can TL 9000 Contribute to Telecom's Turnaround?" *Quality Progress*, February 2004, pp. 38–45.
11. NASA Office of Safety and Mission Assurance, pbma.hq.nasa.gov.
12. Carnegie Mellon University, "Concepts of Operations for the Capability Maturity Model® Integration (CMMI)," August 11, 1999.
13. S. L.K. Briggs, "Next Generation ISO 14001," *Quality Progress*, August 2004, pp. 75–77.
14. M. Block, "ISO 14001 Revision Nears Completion," *Quality Progress*, February 2004.
15. P. Scott, "The Moving Goal Posts from Environmental to Corporate Responsibility," *ISO Management Systems*, September–October 2003, p. 30.

16. U.S. Food and Drug Administration, *Hazard Analysis and Critical Control Point Principles and Application Guidelines*, adopted August 14, 1997.
17. Japanese Union of Scientists and Engineers, "What Is the Deming Prize," www.juse.or.jp/e/deming/.
18. "The Deming Prize and Development of Quality Control/Management in Japan," JUSE website, www.juse.or.jp/e/deming/.
19. Baldrige National Quality Award Program, "About Us," www.nist.gov/baldrige/about/index.cfm.
20. Baldrige National Quality Award Program, *Criteria for Performance Excellence 2011-2012*, www.nist.gov/baldrige/publications/business_nonprofit_criteria.cfm.
21. Baldrige National Quality Award Program, *Criteria for Performance Excellence*, NIST, 2012a, www.baldrige.nist.gov.
22. Baldrige National Quality Award Program, *Fees*, NIST, 2012b, www.nist.gov/baldrige/publications/.../Baldrige_Application_Fees.pdf.
23. A. N. Link and J. T. Scott, *Economic Evaluation of the Baldrige National Quality Program*, NIST Planning Report 11-2, December 2011.
24. Thomson Reuters, *100 Top Hospitals: Study Overview and Research Findings*, 18th ed., Ann Arbor, MI, March 2011.
25. *Quality Progress*, "Baldrige Loses Public Funding," American Society for Quality, January 2012.
26. "Baldrige Expands Reach to Small Business," www.baldrigepe.org/alliance/programs.aspx.
27. *The Shingo Prize for Operational Excellence, Model and Application Guidelines*, Utah State University, May 2012, www.shingoprize.org.
28. P. Wendel, "The European Quality Award and How Texas Instruments Europe Took the Trophy Home," *The Quality Observer*, January 1996.
29. Japan Productivity Center for Socio-Economic Development, "Japan Quality Award," www.jpc-sed.or.jp/eng/award/.
30. A. Starkey, "Continuous Quality Improvement Earns AHCA Recognition," American Health Care Association, August 11, 2003, www.ahca.org/news/nr030711.htm.
31. *Industry Week Magazine*, www.industryweek.com.
32. R. J. Vokurka, G. L. Stading, and J. Brazeal, "A Comparative Analysis of National and Regional Quality Awards," *Quality Progress*, August, 2000.

CHAPTER 24

SAFETY ENGINEERING

Jack B. ReVelle
ReVelle Solutions, LLC
Santa Ana, California

1 INTRODUCTION	692	7.1 Substitution	717
1.1 Background	692	7.2 Isolation	718
1.2 Employee Needs and Expectations	692	7.3 Ventilation	719
2 GOVERNMENT REGULATORY REQUIREMENTS	693	8 DESIGN AND REDESIGN	719
2.1 Environmental Protection Agency	694	8.1 Hardware	719
2.2 Occupational Safety and Health Administration	697	8.2 Process	719
2.3 State-Operated Compliance Programs	698	8.3 Hazardous Material Classification System	720
3 SYSTEM SAFETY	700	8.4 Material Safety Data Sheets	723
3.1 Methods of Analysis	700	8.5 Safety Design Requirements	723
3.2 Fault Tree Technique	702	9 PERSONAL PROTECTIVE EQUIPMENT	724
3.3 Criteria for Preparation/Review of System Safety Procedures	702	9.1 Background	724
3.4 Risk Assessment Process	707	9.2 Planning and Implementing the Use of Protective Equipment	725
4 HUMAN FACTORS ENGINEERING/ ERGONOMICS	708	9.3 Adequacy, Maintenance, and Sanitation	726
4.1 Human–Machine Relationships	708	10 MANAGING THE SAFETY FUNCTION	727
4.2 Human Factors Engineering Principles	709	10.1 Supervisor’s Role	727
4.3 General Population Expectations	710	10.2 Elements of Accident Prevention	727
5 ENGINEERING CONTROLS FOR MACHINE TOOLS	710	10.3 Management Principles	728
5.1 Basic Concerns	710	10.4 Eliminating Unsafe Conditions	729
5.2 General Requirements	712	10.5 Unsafe Conditions Involving Mechanical or Physical Facilities	734
5.3 Danger Sources	713	11 SAFETY TRAINING	735
6 MACHINE SAFEGUARDING METHODS	714	11.1 Specialized Courses	735
6.1 General Classifications	714	11.2 Job Hazard Analysis Training	742
6.2 Guards, Devices, and Feeding and Ejection Methods	715	11.3 Management’s Overview of Training	744
7 ALTERNATIVES TO ENGINEERING CONTROLS	715	11.4 Sources and Types of Training Materials	745
		BIBLIOGRAPHY	745

1 INTRODUCTION

1.1 Background

More than ever before, engineers are aware of and concerned with employee safety and health. The necessity for this involvement was accelerated with the passage of the OSHAct in 1970, but much of what has occurred since that time would have happened whether or not the OSHAct had become the law.

As workplace environments become more technologically complex, the necessity for protecting the workforce from safety and health hazards continues. Typical workplace operations from which workers should be protected are presented in Table 1. Whether workers should be protected through the use of personal protective equipment (PPE), engineering controls, administrative controls, or a combination of these approaches, one fact is clear: It makes good sense to ensure that they receive the most cost-effective protection available. Arguments in support of engineering controls over PPE and vice versa are found everywhere in the current literature. Some of the most persuasive discussions are included in this chapter.

1.2 Employee Needs and Expectations

In 1981 ReVelle and Boulton asked the question, “Who cares about the safety of the worker on the job?” in their award-winning two-part article in *Professional Safety*, “Worker Attitudes and Perceptions of Safety.” The purpose of their study was to learn about worker attitudes and perceptions of safety. To accomplish this objective, they established the following working definition:

Worker Attitudes and Perceptions. As a result of continuing observation, an awareness is developed, as is a tendency to behave in a particular way regarding safety.

To learn about these beliefs and behaviors, they inquired about the following:

1. Do workers think about safety?
2. What do they think about safety in regard to:
 - a. Government involvement in their workplace safety
 - b. Company practices in training and hazard prevention

Table 1 Operations Requiring Engineering Controls and/or PPE

Acidic/basic process and treatments	Grinding
Biological agent processes and treatments	Hoisting
Blasting	Jointing
Boiler/pressure vessel usage	Machinery (mills, lathes, presses)
Burning	Mixing
Casting	Painting
Chemical agent processes and treatments	Radioactive source processes and treatments
Climbing	Sanding
Compressed air/gas usage	Sawing
Cutting	Shearing
Digging	Soldering
Drilling	Spraying
Electrical/electronic assembly and fabrication	Toxic vapor, gas, and mists and dust exposure
Electrical tool usage	Welding
Flammable/combustible/toxic liquid usage	Woodworking

- c. Management attitudes as perceived by the workers
 - d. Co-workers' concern for themselves and others
 - e. Their own safety on the job
3. What do workers think should be done, and by whom, to improve safety in their workplace? The major findings of the ReVelle–Boulton study are summarized here* :
- Half the workers think that government involvement in workplace safety is about right; almost one-fourth think more intervention is needed in such areas as more frequent inspections, stricter regulations, monitoring, and control.
 - Workers in large companies expect more from their employers in providing a safe workplace than workers in small companies. Specifically, they want better safety programs, more safety training, better equipment and maintenance of equipment, more safety inspections and enforcement of safety regulations, and provision of more PPE.
 - Supervisors who talk to their employees about safety and are perceived by them to be serious are also seen as being alert for safety hazards and representative of their company's attitude.
 - Co-workers are perceived by other employees to care for their own safety and for the safety of others.
 - Only 20% of the surveyed workers consider themselves to have received adequate safety training. But more than three-fourths of them feel comfortable with their knowledge to protect themselves on the job.
 - Men are almost twice as likely to wear needed PPE as women.
 - Half the individuals responding said they would correct a hazardous condition if they saw it.
 - Employees who have had no safety training experience almost twice as many on-the-job accidents as their fellow workers who have received such training.
 - Workers who experienced accidents were generally candid and analytical in accepting responsibility for their part in the accident; and 85% said their accidents could have been prevented.

The remainder of this chapter addresses those topics and provides the information that engineering practitioners require to professionally perform their responsibilities with respect to the safety of the workforce.

2 GOVERNMENT REGULATORY REQUIREMENTS[†]

Two agencies of the federal government enforce multiple laws that impact many of the operational and financial decisions of American businesses, large and small. The U.S. Environmental Protection Agency (EPA) has responsibility for administering (listed chronologically by year of passage of each law) the National Environmental Policy Act (NEPA, 1969); the Clean Water Act (CWA, 1970); the Resource Conservation and Recovery Act (RCRA, 1976); the Toxic

* Reprinted with permission from the January 1982 issue of *Professional Safety*, official publication of the American Society of Safety Engineers.

[†] From “Engineering Controls: A Comprehensive Overview” by Jack B. ReVelle. Used by permission of The Merritt Company, Publisher, from T. S. Ferry, *Safety Management Planning*. Copyright © 1982, The Merritt Company, Santa Monica, CA.

Substances Control Act (TSCA, 1976); the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA, aka Superfund, 1980); the Superfund Amendments and Reauthorization Act (SARA, 1986); the Clean Air Act (CAA, 1990); the Oil Pollution Act (OPA, 1990); the Pollution Protection Act (1990); and the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA, 1996). The Occupational Safety and Health Act, also known as OSHA (1970), is enforced by the Occupational Safety and Health Administration (OSHA), a part of the U.S. Department of Labor.

This section addresses the regulatory demands of these federal statutes from the perspective of whether to install engineering controls that would enable companies to meet these standards or simply to discontinue certain operations altogether, that is, whether they can justify the associated costs of regulatory compliance.

2.1 Environmental Protection Agency

National Environmental Policy Act (NEPA, 1969)

This was one of the first acts ever written that establishes the broad national framework for protecting our environment. NEPA's basic policy is to assure that all branches of government give proper consideration to the environment prior to undertaking any major federal action that significantly affects the environment. NEPA requirements are invoked when airports, buildings, military complexes, highways, parkland purchases, and other federal activities are proposed. Environmental assessments (EAs) and environmental impact statements (EISs), which are assessments of the likelihood of impacts from alternative courses of action, are required from all federal agencies and are the most visible NEPA requirements.

Clean Water Act (CWA, 1970)

Growing public awareness and concern for controlling water pollution led to enactment of the Federal Water Pollution Control Act Amendments of 1972. As amended in 1977, this law became commonly known as the Clean Water Act. The CWA established the basic structure for regulating discharges of pollutants into the waters of the United States and gave the EPA the authority to implement pollution control programs such as setting wastewater standards for industry. The act also continued the requirement to set water quality standards for all contaminants in surface waters. It made it unlawful for any person to discharge any pollutant from a point source into navigable waters unless a permit was obtained under its provisions. It also funded the construction of sewage treatment plants under the construction grants programs and recognized the need for planning to address the critical problems posed by nonpoint source pollution.

Subsequent enactments modified some of the earlier CWA provisions. Revisions in 1981 streamlined the municipal construction grants process, improving the capabilities of treatment plants built under the program. Changes in 1987 phased out the construction grants program, replacing it with the State Water Pollution Control Revolving Fund, more commonly known as the Clean Water State Revolving Fund. This new funding strategy addressed water quality needs by building on EPA–state partnerships.

An electronic version of the Clean Water Act, as amended through the enactment of the Great Lakes Legacy Act (GLLA, 2002), is available on the Internet at www.epa.gov/region5/water/cwa.htm. This electronic version annotates the sections of the act with the corresponding sections of the U.S. Code and footnote commentary on the effect of other laws on the current form of the Clean Water Act.

Resource Conservation and Recovery Act (RCRA, 1976)

Enacted in 1976 as an amendment to the Solid Waste Disposal Act, the RCRA sets up a “cradle-to-grave” regulatory mechanism that operates as a tracking system for such wastes

from the moment they are generated to their final disposal in an environmentally safe manner. The act charges the EPA with the development of criteria for identifying hazardous wastes, creating a manifest system for tracking wastes through final disposal, and setting up a permit system based on performance and management standards for generators, transporters, owners, and operators of waste treatment, storage, and disposal facilities. The RCRA is a strong force for innovation that has led to a broad rethinking of chemical processes, that is, to looking at hazardous waste disposal, not just in terms of immediate costs but also with respect to life-cycle costs.

Under the RCRA, wastes are separated into two broad categories: hazardous and non-hazardous. Hazardous wastes are regulated under Subtitle C, and nonhazardous wastes are regulated under Subtitle D. RCRA Subtitle D assists waste management officials in developing and encouraging environmentally sound methods for nonhazardous solid waste disposal. "Solid waste" is a broad term. It is not based on the physical form of the material, but on whether the material is a waste.

The Federal Hazardous and Solid Waste Amendments (HSWA, 1984) to the RCRA required phasing out land disposal of hazardous waste. Some other mandates of this strict law include increased enforcement authority for the EPA, more stringent hazardous waste management standards, and a comprehensive underground storage tank program.

Toxic Substances Control Act (TSCA, 1976)

Until the TSCA, the federal government was not empowered to prevent chemical hazards to health and the environment by banning or limiting chemical substances at a germinal, premarket stage. Through this act, production workers, consumers, indeed all Americans, are protected by an equitably administered early warning system controlled by the EPA. This broad law authorizes the EPA administrator to issue rules to prohibit or limit the manufacturing, processing, or distribution of any chemical substance or mixture that "may present an unreasonable risk of injury to health or the environment." The EPA administrator may require testing—at a manufacturer's or processor's expense—of a substance after finding that:

- The substance may present an unreasonable risk to health or the environment.
- There may be a substantial human or environmental exposure to the substance.
- Insufficient data and experience exist for judging a substance's health and environmental effects.
- Testing is necessary to develop such data.

This legislation is designed to cope with hazardous chemicals like kepone, vinyl chloride, asbestos, fluorocarbon compounds (Freons), and polychlorinated biphenyls (PCBs).

Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA, aka Superfund, 1980)

This act provides a federal "superfund" to clean up uncontrolled or abandoned hazardous waste sites as well as accidents, spills, and other emergency releases of pollutants and contaminants into the environment. Through CERCLA, the EPA was given power to seek out those parties responsible for any release and assure their cooperation in the cleanup.

The EPA cleans up orphan sites when potentially responsible parties cannot be identified or located or when they fail to act. Through various enforcement tools, the EPA obtains private party cleanup through orders, consent decrees, and other small party settlements. The EPA also recovers costs from financially viable individuals and companies once a response action has been completed.

The EPA is authorized to implement the act in all 50 states and U.S. territories. Superfund site identification, monitoring, and response activities in states are coordinated through the state environmental protection or waste management agencies.

Superfund Amendments and Reauthorization Act (SARA, 1986)

This act amended the CERCLA, or “Superfund.” SARA reflected the EPA’s experience in administering the complex Superfund program during its first six years and made several important changes and additions to the program:

- It stressed the importance of permanent remedies and innovative treatment technologies in cleaning up hazardous waste sites.
- It required Superfund actions to consider the standards and requirements found in other state and federal environmental laws and regulations.
- It provided new enforcement authorities and settlement tools.
- It increased state involvement in every phase of the Superfund program.
- It increased the focus on human health problems posed by hazardous waste sites.
- It encouraged greater citizen participation in making decisions on how sites should be cleaned up.
- It increased the size of the trust fund to \$8.5 billion.

SARA also required the EPA to revise the Hazard Ranking System (HRS) to ensure that it accurately assessed the relative degree of risk to human health and the environment posed by uncontrolled hazardous waste sites that may be placed on the National Priorities List (NPL).

Clean Air Act (CAA, 1990)

Although this act is a federal law covering the entire country, the states do much of the work to carry it out. For example, a state air pollution agency holds a hearing on a permit application by a power or chemical plant or fines a company for violating air pollution limits. Under this law, the EPA sets limits on how much of a pollutant can be in the air anywhere in the United States. This ensures that all Americans have the same basic health and environmental protections. The law allows individual states to have stronger pollution controls, but states are not allowed to have weaker pollution controls than those set for the entire country. The law recognizes that it makes sense for states to take the lead in carrying out the CAA because pollution control problems often require special understanding of local industries, geography, housing patterns, etc.

States must develop state implementation plans (SIPs) that explain how each state will do its job under the CAA. A SIP is a collection of the regulations that a state will use to clean up polluted areas. The states must involve the public, through hearings and opportunities to comment, in the development of each SIP. The EPA must approve each SIP, and if a SIP is not acceptable, the EPA can take over enforcing the CAA in that state. The U.S. government, through the EPA, assists the states by providing scientific research, expert studies, engineering designs, and money to support clean air programs.

Oil Pollution Act (OPA, 1990)

The OPA streamlined and strengthened the EPA’s ability to prevent and respond to catastrophic oil spills. A trust fund financed by a federal tax on oil is available to clean up spills when the responsible party is incapable or unwilling to do so. The OPA requires oil storage facilities and vessels to submit to the federal government plans detailing how they will respond to large discharges. The EPA has published regulations for above-ground storage facilities while the U.S. Coast Guard has done the same for oil tankers. The OPA also requires the development of area contingency plans (ACPs) to prepare and plan for oil spill response on a regional basis.

Pollution Protection Act (PPA, 1990)

The PPA focused industry, government, and public attention on reducing the amount of pollution through cost-effective changes in production, operation, and raw materials use. Opportunities for source reduction are often not realized because of existing regulations and the industrial resources required for compliance, focus on treatment, and disposal. Source reduction is fundamentally different and more desirable than waste management or pollution control.

Pollution prevention also includes other practices that increase efficiency in the use of energy, water, or other natural resources and protect our resource base through conservation. Practices include recycling, source reduction, and sustainable agriculture.

Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA, 1996)

The primary focus of the FIFRA was to provide federal control of pesticide distribution, sale, and use. The EPA was given the authority under FIFRA not only to study the consequences of pesticide usage but also to require users (farmers, utility companies, and others) to register when purchasing pesticides. Through later amendments to the law, users must also take exams for certification as applicators of pesticides. All pesticides use in the United States must be registered (licensed) by the EPA. Registration assures that pesticides will be properly labeled and that, if in accordance with the specifications, will not cause unreasonable harm to the environment.

2.2 Occupational Safety and Health Administration*

The Occupational Safety and Health Act (OSHAct), a federal law that became effective on April 28, 1971, is intended to pull together all federal and state occupational safety and health enforcement efforts under a federal program designed to establish uniform codes, standards, and regulations. The expressed purpose of the act is “to assure, as far as possible, every working woman and man in the Nation safe and healthful working conditions, and to preserve our human resources.” To accomplish this purpose, the promulgation and enforcement of safety and health standards is provided for, as well as research, information, education, and training in occupational safety and health. Perhaps no single piece of federal legislation has been more praised and, conversely, more criticized than the OSHAct, which basically is a law requiring virtually all employers to ensure that their operations are free of hazards to workers.

Occupational Safety and Health Standards

When Congress passed the OSHAct of 1970, it authorized the promulgation, without further public comment or hearings, of groups of already codified standards. The initial set of standards of the act (Part 1910, published in the *Federal Register* on May 29, 1971) thus consisted in part of standards that already had the force of law, such as those issued by authority of the Walsh-Healey Act, the Construction Safety Act, and the 1958 amendments to the Longshoremen’s and Harbor Workers’ Compensation Act. A great number of the adopted standards, however, derived from voluntary national consensus standards previously prepared by groups such as the American National Standards Institute (ANSI) and the National Fire Protection Association (NFPA).

The OSHAct defines the term “occupational safety and health standard” as meaning “a standard which requires conditions or the adoption or use of one or more practices, means, methods, operations or processes, reasonably necessary or appropriate to provide safe or healthful employment and places of employment.” Standards contained in Part 1910[†] are applicable

* From R. De Reamer, *Modern Safety and Health Technology*. Copyright © 1980. Reprinted by permission of John Wiley & Sons, New York.

[†] The Occupational Safety and Health Standards, Title 29, CFR Chapter XVIII. Parts 1910, 1926, and 1915–1918 are available at all OSHA regional and area offices.

to general industry; those contained in Part 1926 are applicable to the construction industry; and standards applicable to ship repairing, shipbuilding, and longshoring are contained in Parts 1915–1918. These OSHA standards fall into the following four categories, with examples for each type:

1. *Specification Standards*. Standards that give specific proportions, locations, and warning symbols for signs that must be displayed.
2. *Performance Standards*. Standards that require achievement of, or within, specific minimum or maximum criteria.
3. *Particular Standards (Vertical)*. Standards that apply to particular industries, with specifications that relate to the individual operations.
4. *General Standards (Horizontal)*. Standards that can apply to any workplace and relate to broad areas (environmental control, walking surfaces, exits, illumination, etc.).

OSHA is authorized to promulgate, modify, or revoke occupational safety and health standards. It also has the authority to promulgate emergency temporary standards where it is found that employees are exposed to grave danger. Emergency temporary standards can take effect immediately on publication in the *Federal Register*. Such standards remain in effect until superseded by a standard promulgated under the procedures prescribed by the OSHAct—notice of proposed rule in the *Federal Register*; invitation to interested persons to submit their views, and a public hearing if required.

Required Notices and Records

During an inspection the compliance officer will ascertain whether the employer has:

- Posted notice informing employees of their rights under the OSHAct (Job Safety and Health Protection, OSHAct poster).
- Maintained a log of recordable injuries and illnesses (OSHA Form No. 300, Log and Summary of Occupational Injuries and Illnesses).
- Maintained the Supplementary Record of Occupational Injuries and Illnesses (OSHA Form No. 301).
- Annually posted the Summary of Occupational Injuries and Illnesses (OSHA Form No. 200). This form must be posted no later than February 1 and must remain in place until March 1.
- Made a copy of the OSHAct and OSHA safety and health standards available to employees on request.
- Posted boiler inspection certificates, boiler licenses, elevator inspection certificates, and so on.

2.3 State-Operated Compliance Programs

The OSHAct encourages each state to assume the fullest responsibility for the administration and enforcement of occupational safety and health programs. For example, federal law permits any state to assert jurisdiction, under state law, over any occupational or health standard not covered by a federal standard. Section 18 of the Occupational Safety and Health Act of 1970 (OSHAct) encourages states to develop and operate their own job safety and health programs. OSHA approves and monitors state plans and provides 50% of an approved plan's operating costs. States must set job safety and health standards that are "at least as effective as" comparable federal standards. (Most states adopt standards identical to federal ones.) States have the option to promulgate standards covering hazards not addressed by federal standards.

A state must conduct inspections to enforce its standards, cover public (state and local government) employees, and operate occupational safety and health training and education programs. In addition, most states provide free, on-site consultation to help employers identify and correct workplace hazards. Such consultation may be provided either under the plan or through a special agreement under Section 21(d) of the OSHA Act.

In addition, any state may assume responsibility for the development and enforcement of its own occupational safety and health standards for those areas now covered by federal standards. However, the state must first submit a plan for approval by the Labor Department's Occupational Safety and Health Administration. Many states have done so. To gain OSHA approval for a *developmental plan*—the first step in the state plan process—a state must assure OSHA that within three years it will have in place all the structural elements necessary for an effective occupational safety and health program. These elements include appropriate legislation; regulations and procedures for standards setting, enforcement, appeal of citations, and penalties; and a sufficient number of qualified enforcement personnel.

Once a state has completed and documented all its developmental steps, it is eligible for *certification*. Certification renders no judgment as to actual state performance but merely attests to the structural completeness of the plan.

At any time after initial plan approval, when it appears that the state is capable of independently enforcing standards, OSHA may enter into an "*operational status agreement*" with the state. This commits OSHA to suspend the exercise of discretionary federal enforcement in all of certain activities covered by the state.

The ultimate accreditation of a state's plan is called *final approval*. When OSHA grants final approval to a state under Section 18(e) of the OSHA Act, it relinquishes its authority to cover occupational safety and health matters covered by the state. After at least one year following certification, the state becomes eligible for approval if OSHA determines that the state is providing, in actual operation, worker protection "at least as effective as" the protection provided by the federal program. The state must also meet 100% of the established compliance staffing levels (benchmarks) and participate in OSHA's computerized inspection data system before OSHA can grant final approval.

Employees finding workplace safety and health hazards may file a formal complaint with the appropriate state or with the appropriate OSHA regional administrator. Complaints will be investigated and should include the name of the workplace, type(s) of hazard(s) observed and any other pertinent information. Anyone finding inadequacies or other problems in the administration of a state's program may file a Complaint About State Program Administration (CASPA) with the appropriate OSHA regional administrator as well. A complainant's name is kept confidential. OSHA investigates all such complaints and, where complaints are found to be valid, requires appropriate corrective action on the part of the state.

Certain states are now operating under an approved state plan. These states may have adopted the existing federal standards or may have developed their own standards. Some states also have changed the required poster. Individuals need to know whether they are covered by an OSHA-approved state plan operation or are subject to the federal program to determine which set of standards and regulations (federal or state) apply to you. The easiest way to determine this is to call the nearest OSHA area office. If you are subject to state enforcement, the OSHA area office will explain this, explain whether the state is using the federal standards, and provide you with information on the poster and on the OSHA recordkeeping requirements. After that, the OSHA area office will refer you to the appropriate state government office for further assistance. This assistance also may include free on-site consultation visits. If you are subject to state enforcement, you should take advantage of this service.

There are currently 22 states and jurisdictions operating complete state plans (covering the private sector as well as state and local government employees) and 4—Connecticut, New Jersey, New York, and the Virgin Islands—that cover public employees only. (Eight other states

Table 2 States Operating Under OSHA-Approved Plans as of April 16, 2004

1. Alaska	14. New Jersey ^a
2. Arizona	15. New York ^a
3. California	16. North Carolina
4. Connecticut ^a	17. Oregon
5. Hawaii	18. Puerto Rico
6. Indiana	19. South Carolina
7. Iowa	20. Tennessee
8. Kentucky	21. Utah
9. Maryland	22. Vermont
10. Michigan	23. Virginia
11. Minnesota	24. Virgin Islands ^a
12. Nevada	25. Washington
13. New Mexico	26. Wyoming

^aThese plans cover public sector employees only.

were approved at one time but subsequently withdrew their programs.) Table 2 lists those states operating under OSHA-approved state plans as of April 16, 2004.

3 SYSTEM SAFETY*

System safety is when situations having accident potential are examined in a step-by-step cause-effect manner, tracing a logical progression of events from start to finish. System safety techniques can provide meaningful predictions of the frequency and severity of accidents. However, their greatest asset is the ability to identify many accident situations in the system that would have been missed if less detailed methods had been used.

3.1 Methods of Analysis

A system cannot be understood simply in terms of its individual elements or component parts. If an operation of a system is to be effective, all parts must interact in a predictable and a measurable manner within specific performance limits and operational design constraints.

In analyzing any system, three basic components must be considered: (1) the equipment (or machines); (2) the operators and supporting personnel (maintenance technicians, material handlers, inspectors, etc.); and (3) the environment in which both workers and machines are performing their assigned functions. Several analysis methods are available:

- *Gross-Hazard Analysis*. Performed early in design; considers overall system as well as individual components; it is called “gross” because it is the initial safety study undertaken.
- *Classification of Hazards*. Identifies types of hazards disclosed in the gross-hazard analysis and classifies them according to potential severity (Would defect or failure be catastrophic?); indicates actions and/or precautions necessary to reduce hazards. May involve preparation of manuals and training procedures.
- *Failure Modes and Effects*. Considers kinds of failures that might occur and their effect on the overall product or system. Example: effect on system that will result from failure of single component (e.g., a resistor or hydraulic valve).

* From R. De Reamer, *Modern Safety and Health Technology*. Copyright © 1980. Reprinted by permission of John Wiley & Sons, New York.

- *Hazard Criticality Ranking*. Determines statistical, or quantitative, probability of hazard occurrence; ranking of hazards in the order of “most critical” to “least critical.”
- *Fault Tree Analysis*. Traces probable hazard progression. *Example*: If failure occurs in one component or part of the system, will fire result? Will it cause a failure in some other component?
- *Energy Transfer Analysis*. Determines interchange of energy that occurs during a catastrophic accident or failure. Analysis is based on the various energy inputs to the product or system and how these inputs will react in the event of failure or catastrophic accident.
- *Catastrophe Analysis*. Identifies failure modes that would create a catastrophic accident.
- *System–Subsystem Integration*. Involves detailed analysis of interfaces, primarily between systems.
- *Maintenance Hazard Analysis*. Evaluates performance of the system from a maintenance standpoint. Will it be hazardous to service and maintain? Will maintenance procedures be apt to create new hazards in the system?
- *Human Error Analysis*. Defines skills required for operation and maintenance. Considers failure modes initiated by human error and how they would affect the system. The question of whether special training is necessary should be a major consideration in each step.
- *Transportation Hazard Analysis*. Determines hazards to shippers, handlers, and bystanders. Also considers what hazards may be “created” in the system during shipping and handling.

Other quantitative methods have successfully been used to recommend a decision to adopt engineering controls, PPE, or some combination. Some of these methods are as follows* :

Expected Outcome Approach. Since safety alternatives involve accident costs that occur more or less randomly according to probabilities that might be estimated, a valuable way to perform needed economic analyses for such alternatives is to calculate expected outcomes.

- *Decision Analysis Approach*. A recent extension of systems analysis, this approach provides useful techniques for transforming complex decision problems into a sequentially oriented series of smaller, simpler problems. This means that decision-makers can select reasoned choices that will be consistent with their perceptions about the uncertainties involved in a particular problem together with their fundamental attitudes toward risk taking.
- *Mathematical Modeling*. Usually identified as an “operations research” approach, numerous mathematical models have demonstrated potential for providing powerful analysis insights into safety problems. These include dynamic programming, inventory-type modeling, linear programming, queue-type modeling, and Monte Carlo simulation.

There is a growing body of literature about these formal analytical methods and others not mentioned in this chapter, including failure mode and effect (FME), technique for human error prediction (THERP), system safety hazard analysis, and management oversight and risk tree (MORT). All have their place. Each to a greater or lesser extent provides a means of overcoming the limitations of intuitive, trial-and-error analysis.

* From J. B. ReVelle, *Engineering Controls: A Comprehensive Overview*. Used by permission of The Merritt Company, Publisher, from T. S. Ferry, *Safety Management Planning*. Copyright © 1982, The Merritt Company, Santa Monica, CA.

Regardless of the method or methods used, the systems concept of hazard recognition and analysis makes available a powerful tool of proven effectiveness for decision making about the acceptability of risks. To cope with the complex safety problems of today and the future, engineers must make greater use of system safety techniques.

3.2 Fault Tree Technique*

When a problem can be stated quantitatively, management can assess the risk and determine the trade-off requirements between risk and capital outlay. Structuring key safety problems or vital decision making in the form of fault paths can greatly increase communication of data and subjective reasoning. This technique is called fault tree analysis. The transferability of data among management, engineering staff, and safety personnel is a vital step forward.

Another important aspect of this system safety technique is a phenomenon that engineers have long been aware of in electrical networks. That is, an end system formed by connecting several subsystems is likely to have entirely different characteristics from any of the subsystems considered alone. To fully evaluate and understand the entire system's performance with key paths of potential failure, the engineer must look at the entire system. Only then can he or she look meaningfully at each of the subsystems.

Figure 1 introduces the symbols most commonly used in fault tree analysis.

3.3 Criteria for Preparation/Review of System Safety Procedures[†]

Correlation between Procedure and Hardware

1. Is there a statement of hardware configuration to which the procedure was written?
2. Has background descriptive or explanatory information been provided where needed?
3. Does the hardware reflect or reference the latest revisions of drawings, manuals, or other procedures?

Adequacy of the Procedure

1. Is the procedure the best way to do the job?
2. Is the procedure easy to understand?
3. Is the detail appropriate—not too much, not too little?
4. Is it clear, concise, and free from ambiguity that could lead to wrong decisions?
5. Are calibration requirements clearly defined?
6. Are critical red-line parameters identified and clearly defined? Are required values specified?
7. Are corrective controls of the above parameters clearly defined?
8. Are all values, switches, and other controls identified and defined?
9. Are pressure limits, caution notes, safety distances, or hazards peculiar to this operation clearly defined?
10. Are hard-to-locate components adequately defined and located?
11. Are jigs and arrangements provided to minimize error?

* From R. De Reamer, *Modern Safety and Health Technology*. Copyright © 1980. Reprinted by permission of John Wiley & Sons, New York.

[†] Reprinted from *MORT Safety Assurance Systems*, pp. 278–283, by courtesy of Marcel Dekker, New York.

12. Are job safety requirements defined, e.g., power off, pressure down, and have tools been checked for sufficiency?
13. Is the system operative at end of job?
14. Has the hardware been evaluated for human factors and behavioral stereotype problems? If not corrected, are any such clearly identified?
15. Have monitoring points and methods of verifying adherence been specified?

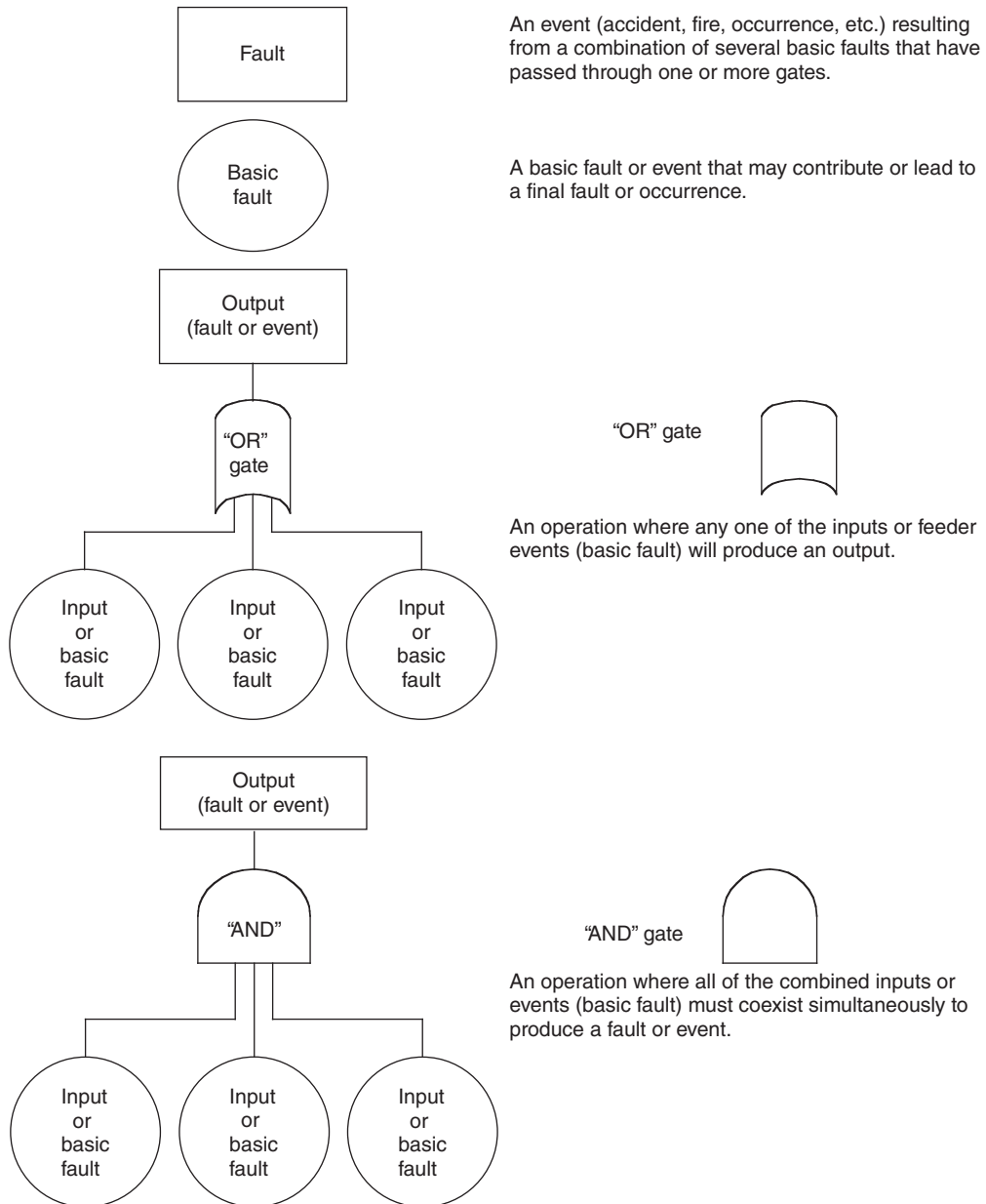


Figure 1 Most common symbols used in fault tree analysis.

16. Are maintenance and/or inspection to be verified? If so, is a log provided?

An event (accident, fire, occurrence, etc.) resulting from a combination of several basic faults that have passed through one or more gates.

A basic fault or event that may contribute or lead to a final fault or occurrence.

“OR” gate

An operation where any one of the inputs or feeder events (basic fault) will produce an output.

“AND” gate

An operation where all of the combined inputs or events (basic fault) must coexist simultaneously to produce a fault or event.

17. Is safe placement of process personnel or equipment specified?

18. Have errors in previous, similar processes been studied for cause? Does this procedure correct such causes?

Accuracy of the Procedure

1. Has the capacity to accomplish the specified purpose been verified by internal review?
2. Have all gauges, controls, valves, etc., been called out, described, and labeled exactly as they actually are?
3. Are all setpoints or other critical controls, etc., compatible with values in control documents?
4. Are the safety limitations adequate for job to be performed?
5. Are all steps in the proper sequence?

Adequacy and Accuracy of Supporting Documentation

1. Are all necessary supporting drawings, manuals, data sheets, sketches, etc., either listed or attached?
2. Are all interfacing procedures listed?

Securing Provisions

1. Are there adequate instructions to return the facility or hardware to a safe operating or standby condition?
2. Do securing instructions provide step-by-step operations?

Backout Provisions

1. Can the procedure put any component or system in a condition that could be dangerous?
2. If so, does the procedure contain emergency shutdown or backout procedures either in an appendix or as an integral part?
3. Is a backout procedure (or instructions for its use) included at proper place?

Emergency Measures

1. What are the procedures for action in case of emergency conditions?
2. Does the procedure involve critical actions such that preperformance briefing on possible hazards is required?

3. Are adequate instructions either included or available for action to be taken under emergency conditions? Are they in the right place?
4. Are adequate shutdown procedures available? Are coverall systems involved and available for emergency reentry teams?
5. Are requirements for the emergency team for accident recovery, troubleshooting, or investigative purposes specified where necessary? Are conditions under which the emergency team will be used described? Are the hazards they may encounter or must avoid identified?
6. Does the procedure consider interfaces in shutdown procedures?
7. How will changes be handled? What are the thresholds for changes requiring review?
8. Have the emergency procedures been tested under a range of conditions that may be encountered, e.g., at night during power failure?

Caution and Warning Notes

1. Are caution and warning notes included where appropriate?
2. Do caution and warning notes precede operational steps containing potential hazards?
3. Are they adequate to describe the potential hazard?
4. Are major cautions and warnings called out in the general introduction as well as prior to steps?
5. Do notes appear as separate entries with distinctive bold type or other emphatic display?
6. Do they include supporting safety control (health physics, safety engineer, etc.) if needed at specific required steps in the procedure?

Requirements for Communications and Instrumentation

1. Are adequate means of communication provided?
2. Will loss of communications create a hazard?
3. Is a course of action clearly defined for loss of required communications?
4. Is verification of critical communication included prior to point of need?
5. Will loss of control or monitoring capability of critical functions create a hazard to people or hardware?
6. Are alternative means, or a course of action to regain control or monitoring functions, clearly defined?
7. Are the above situations flagged by cautions and warnings?

Sequence-of-Events Considerations

1. Can any operation initiate an unscheduled or out-of-sequence event?
2. Could it induce a hazardous condition?
3. Are these operations identified by warnings or cautions?
4. Are they covered by emergency shutdown and backout procedures?
5. Are all steps sequenced properly? Will the sequence contribute to or create a hazard?
6. Are all steps that, if performed out of sequence, could cause a hazard identified and flagged?

7. Have all noncompatible simultaneous operations been identified and suitably restricted?
8. Have these been prohibited by positive callout or separation in step-by-step inclusion within the text of the procedure?

Environmental Considerations (Natural or Induced)

1. Have the environmental requirements been specified that constrain the initiation of the procedure or require shutdown or evacuation, once in progress?
2. Have the induced environments (toxic or explosive atmospheres, etc.) been considered?
3. Have all latent hazards (pressure, height, voltage, etc.) in adjacent environments been considered?
4. Are there induced hazards from simultaneous performance of more than one procedure by personnel within a given space?

Personnel Qualification Statements

1. Has the requirement for certified personnel been considered?
2. Has the required frequency of recheck of personnel qualifications been considered?

Interfacing Hardware and Procedures Noted

1. Are all interfaces described by detailed callout?
2. Are interfacing operating procedures identified or written to provide ready equipment?
3. Where more than one organizational element is involved, are proper liaison and areas of responsibility established?

Procedure Sign-Off

1. Is the procedure to be used as an in-hand, literal checklist?
2. Have the step sign-off requirements been considered and identified and appropriate spaces provided in the procedure?
3. Have the procedure completion sign-off requirements been indicated (signature, authority, date, etc.)?
4. Is supervisor verification of correct performance required?

General Requirements

1. Does the procedure discourage a shift change during performance or accommodate a shift change?
2. Where shift changes are necessary, do they include or reference shift overlap and briefing requirements?
3. Are mandatory inspection, verification, and system validation required whenever the procedure requires breaking into and reconnecting a system?
4. Are safety prerequisites defined? Are all safety instructions spelled out in detail to all personnel?
5. Does the procedure require prechecks of supporting equipment to ensure compatibility and availability?

6. Is consideration for unique operations written in?
7. Does the procedure require walk-through or talk-through dry runs?
8. What are the general supervision requirements, e.g., what is the protocol for transfer of supervisor responsibilities to a successor?
9. Are the responsibilities of higher supervision specified?

Reference Considerations

1. Have applicable quality assurance and reliability standards been considered?
2. Have applicable codes, standards, and regulations been considered?
3. Does the procedure comply with control documents?
4. Have hazards and system safety degradations been identified and considered against specific control manuals, standards, and procedures?
5. Have specific prerequisite administrative and management approvals been complied with?
6. Have comments been received from the people who will do the work?

Special Considerations

1. Has a documented safety analysis been considered for safety-related deviations from normal practices or for unusual or unpracticed maneuvers'?
2. Have new restrictions or controls become effective that affect the procedure in such a manner that new safety analyses may be required?

3.4 Risk Assessment Process

*Risk Score Formula**

William T. Fine is credited with having developed a unique procedure that responds to potential accident hazards by mathematically determining organizational priorities for action. His methodology assigns weight to known controlling factors and then determines the overall risk for a specific hazardous situation by calculating a "risk score" that indicates the relative urgency of the corresponding remedial action.

The risk score formula is

$$\text{Risk score} = \text{consequences} \times \text{exposure} \times \text{probability}$$

In using the formula, the numeric ratings or weights are subjectively assigned to each factor based on the judgment and experience of the investigator(s) making the calculation. The following is a review of the formula:

- *Consequences.* The most probable results of an accident due to a hazard being considered, including both injuries and property damage. Numerical ratings are assigned for the most likely consequences of the accident starting from 100 points for a catastrophe down through various degrees of severity to 1 point for a minor cut or bruise.
- *Exposure.* The frequency of occurrence of the *hazard event*, which is the *first undesired event* that *could* start the accident sequence. The frequency with which the hazard event is most likely to occur is rated from *continuous occurrence* with 10 points through various lesser degrees of exposure down to 0.5 point for *extremely remote*.

* From W. T. Fine, "Mathematical Evaluations for Controlling Hazards," in *Selected Readings in Safety*, J. Widener (ed.), Academy Press, Macon, GA.

- *Probability*. The likelihood that, once the hazard event occurs, the *complete accident sequence of events will follow* with the timing and coincidence that results in an accident and its consequences. The ratings start at 10 points if the complete accident sequence is *most likely and expected* and continue down to 0.1 point for the “one in a million” or practically impossible chance.

Multiplying the *maximum* points for each factor (consequences = 100, exposure = 10, and probability = 10) results in the *greatest* possible risk score of 10,000. Multiplying the *minimum* points for each factor (consequences = 1, exposure = 0.1, and probability = 0.1) results in the *least* possible risk score of 0.05.

Fine recommended the following risk score–action relationship:

Risk Score	Recommended Action
0.05–89	Hazard should be eliminated without delay, but situation is not an emergency.
90–200	Urgent. Requires attention as soon as possible.
201–10,000	Immediate correction required. Activity should be discontinued until hazard is abated.

For a newly discovered hazard, the risk score–action relationship provides important guidance with respect to necessary action.

4 HUMAN FACTORS ENGINEERING/ERGONOMICS*

4.1 Human–Machine Relationships

- Human factors engineering is defined as “the application of the principles, laws, and quantitative relationships which govern man’s response to external stress to the analysis and design of machines and other engineering structures, so that the operator of such equipment will not be stressed beyond his/her proper limit or the machine forced to operate at less than its full capacity in order for the operator to stay within acceptable limits of human capabilities.”[†]
- A principal objective of the supervisor and safety engineer in the development of safe working conditions is the elimination of bottlenecks, stresses and strains, and psychological booby traps that interfere with the free flow of work. The less operators have to fear from their jobs or machines, the more attention they can give to their work.
- In the development of safe working conditions, attention is given to many things, including machine design and machine guarding, PPE, plant layout, manufacturing methods, lighting, heating, ventilation, removal of air contaminants, and the reduction of noise. Adequate consideration of each of these areas will lead to a proper climate for accident prevention, increased productivity, and worker satisfaction.
- The human factors engineering approach to the solution of the accident problem is to build machines and working areas around the operator, rather than place him or her in

* From R. De Reamer, *Modern Safety and Health Technology*. Copyright © 1980. Reprinted by permission of John Wiley & Sons, New York.

† Courtesy of Theodore F. Hatch, Professor (retired), by permission.

a setting without regard to his or her requirements and capacities. Unless this is done, it is hardly fair to attribute so many accidents to human failure, as is usually the case.

- If this point of view is carried out in practice, fewer accidents should result, training costs should be reduced, and extensive redesign of equipment after it is put into use should be eliminated.
- All possible faults in equipment and in the working area, as well as the capacities of the operator, should be subjected to advance analysis. If defects are present, it is only a matter of time before some operator “fails” and has an accident.
- Obviously, the development of safe working conditions involves procedures that may go beyond the occasional safety appraisal or search for such obvious hazards as an oil spot on the floor, a pallet in the aisle, or an unguarded pinch point on a new lathe.

Human–machine relationships have improved considerably with increased mechanization and automation. Nevertheless, with the decrease in manual labor has come specialization, increased machine speeds, and monotonous repetition of a single task, which create work relationships involving several physiological and psychological stresses and strains. Unless this scheme of things is recognized and dealt with effectively, many real problems in the field of accident prevention may be ignored.

4.2 Human Factors Engineering Principles

- Human factors engineering, or *ergonomics*,* as it is sometimes called, developed as a result of the experience in the use of highly sophisticated equipment in World War II. The ultimate potentialities of complex instruments of war could not be realized because the human operators lacked the necessary capabilities and endurance required to operate them. This discipline now has been extended to many areas. It is used extensively in the aircraft and aerospace industry and in many other industries to achieve more effective integration of humans and machines.
- The analysis should consider all possible faults in the equipment, in the work area, and in the worker, including a survey of the nature of the task, the work surroundings, the location of controls and instruments, and the way the operator performs his or her duties. The questions of importance in the analysis of machines, equipment, processes, plant layout, and the worker will vary with the type and purpose of the operation but usually will include the following (pertaining to the worker)[†] :
 1. What sense organs are used by the operator to receive information? Does he or she move into action at the sound of a buzzer, blink of a light, reading of a dial, verbal order? Does the sound of a starting motor act as a cue?
 2. What sort of discrimination is called for? Does the operator have to distinguish between lights of two different colors, tones of two different pitches, or two dial readings?
 3. What physical response is he or she required to make: Pull a handle? Turn a wheel? Step on a pedal? Push a button?
 4. What overall physical movements are required in the physical response? Do such movements interfere with his or her ability to continue receiving information through

* The term *ergonomics* was coined from the Greek roots *ergon* (work) and *nomos* (law, rule) and is now currently used to deal with the interactions between humans and such environmental elements as atmospheric contaminants, heat, light, sound, and all tools and equipment pertaining to the workplace.

[†] From R. A. McFarland, “Application of Human Factors Engineering to Safety Engineering Problems,” *National Safety Congress Transactions*, 12 (1967), with permission from the National Safety Council.

his or her sense organs? (For example, would pulling a handle obstruct his or her line of vision to a dial he or she is required to watch?) What forces are required (e.g., torque in turning a wheel)?

5. What are the speed and accuracy requirements of the machine? Is the operator required to watch two pointers to a hairline accuracy in a split second? Or is a fairly close approximation sufficient? If a compromise is necessary, which is more essential: speed or accuracy?
 6. What physiological and environmental conditions are likely to be encountered during normal operation of the machine? Are there any unusual temperatures, humidity conditions, crowded workspace, poor ventilation, high noise levels, toxic chemicals, and so on?
- Pertaining to the machine, equipment, and the surrounding area, these key questions should be asked:
 1. Can the hazard be eliminated or isolated by a guard, ventilating equipment, or other device?
 2. Should the hazard be identified by the use of color, warning signs, blinking lights, or alarms?
 3. Should interlocks be used to protect the worker when he or she forgets or makes the wrong move?
 4. Is it necessary to design the machine, the electrical circuit, or the pressure circuit so it will always be fail-safe?
 5. Is there need for standardization?
 6. Is there need for emergency controls and are controls easily identified and accessible?
 7. What unsafe conditions would be created if the proper operating sequence were not followed?

4.3 General Population Expectations*

- The importance of standardization and normal behavior patterns has been recognized in business and industry for many years. A standard tool will more likely be used properly than will a nonstandard one, and standard procedures will more likely be followed.
- People expect things to operate in a certain way and certain conditions to conform to established standards. These general population “expectations”—the way in which the ordinary person will react to a condition or stimulus—must not be ignored or workers will be literally trapped into making mistakes. A list of general population expectations is given in Table 3.

5 ENGINEERING CONTROLS FOR MACHINE TOOLS[†]

5.1 Basic Concerns

Machine tools (such as mills, lathes, shearers, punch presses, grinders, drills, and saws) provide an example of commonplace conditions where only a limited number of items of personal

* From R. De Reamer, *Modern Safety and Health Technology*. Copyright © 1980. Reprinted by permission of John Wiley & Sons, New York.

[†] From J. B. ReVelle, *Engineering Controls: A Comprehensive Overview*. Used by permission of the Merritt Company, Publisher, from T. S. Ferry, *Safety Management Planning*. Copyright © 1982, The Merritt Company, Santa Monica, CA.

Table 3 General Population Expectations

-
1. Doors are expected to be at least 6 ft, 6 in. in height.
 2. The level of the floor at each side of a door is expected to be the same.
 3. Stair risers are expected to be of the same height.
 4. It is a normal pattern for persons to pass to the left on motorways (some countries excluded).
 5. People expect guardrails to be securely anchored.
 6. People expect the hot-water faucet to be on the left side of the sink, the cold-water faucet on the right, and the faucet to turn to the left (counterclockwise) to let the water run and to the right to turn the water off.
 7. People expect floors to be nonslippery.
 8. Flammable solvents are expected to be found in labeled, red containers.
 9. The force required to operate a lever, push a cart, or turn a crank is expected to go unchanged.
 10. Knobs on electrical equipment are expected to turn clockwise for “on,” to increase current, and counterclockwise for “off.”
 11. For control of vehicles in which the operator is riding, the operator expects a control motion to the right or clockwise to result in a similar motion of his or her vehicle and vice versa.
 12. Very large objects or dark objects imply “heaviness.” Small objects or light-colored ones imply “lightness.” Large heavy objects are expected to be “at the bottom.” Small, light objects are expected to be “at the top.”
 13. Seat heights are expected to be at a certain level when a person sits.
-

protective gear are available for use. In such cases as these, the problem to be solved is not PPE versus engineering controls, but rather which engineering control(s) should be used to protect the machine operator. A summary of employee safeguards is given in Table 4. The list of possible machinery-related injuries is presented later in Table 11. There seem to be as many hazards created by moving machine parts as there are types of machines. Safeguards are essential for protecting workers from needless and preventable injuries.

A good rule to remember is: Any machine part, function, or process that may cause injury must be safeguarded. Where the operation of a machine or accidental contact with it can injure the operator or others in the vicinity, the hazard must be either controlled or eliminated.

Dangerous moving parts in these three basic areas need safeguarding:

- *Point of Operation.* That point where work is performed on the material, such as cutting, shaping, boring, or forming of stock.
- *Power Transmission Apparatus.* All components of the mechanical system that transmit energy to the part of the machine performing the work. These components include flywheels, pulleys, belts, connecting rods, couplings, cams, spindles, chains, cranks, and gears.
- *Other Moving Parts.* All parts of the machine that move while the machine is working. These can include reciprocating, rotating, and transverse moving pans as well as feed mechanisms and auxiliary parts of the machine.

A wide variety of mechanical motions and actions may present hazards to the worker. These can include the movement of rotating teeth and any parts that impact or shear. These different types of hazardous mechanical motions and actions are basic to nearly all machines, and recognizing them is the first step toward protecting workers from the danger they present.

Table 4 Summary of Employee Safeguards

To Protect	PPE to Use	Engineering Controls to Use
Breathing	Self-contained breathing apparatus, gas masks, respirators, alarm systems	Ventilation, air filtration systems, critical-level warning systems, electrostatic precipitators
Eyes/face	Safety glasses, filtered lenses, safety goggles, face shield, welding goggles/helmets, hoods	Spark deflectors, machine guards
Feet/legs	Safety boots/shoes, leggings, shin guards	
Hands/arms/body	Gloves, finger cots, jackets, sleeves, aprons, barrier creams	Machine guards, lockout devices, feeding and ejection methods
Head/neck	Bump caps, hard hats, hair nets	Toe boards
Hearing	Ear muffs, ear plugs, ear valves	Noise reduction, isolation by equipment modification/substitution, equipment lubrication/maintenance programs, eliminate/dampen noise sources, reduce compressed air pressure, change operations ^a
Excessively high/low temperatures	Reflective clothing, temperature-controlled clothing	Fans, air conditioning, heating, ventilation, screens, shields, curtains
Overall	Safety belts, lifelines, grounding mats, slap bars	Electrical circuit grounding, polarized plugs/outlets, safety nets

^aExamples of the types of changes that should be considered include:

- Grinding instead of chipping
- Electric tools in place of pneumatic tools
- Pressing instead of forging
- Welding instead of riveting
- Compression riveting over pneumatic riveting
- Mechanical ejection in place of air blast ejection
- Wheels with rubber or composition tires on plant trucks and cars instead of all-metal wheels
- Wood or plastic tote boxes in place of metal tote boxes
- Use of an undercoating on machinery covers.
- Wood in place of all-metal workbenches.

The basic types of hazardous mechanical motions and actions are as follows:

Motions	Actions
Rotating (including in-running nip points)	Cutting Punching
Reciprocating	Shearing
Transverse	Bending

5.2 General Requirements

What must a safeguard do to protect workers against mechanical hazards? Engineering controls must meet these minimum general requirements:

- *Prevent Contact.* The safeguard must prevent hands, arms, or any other part of a worker's body from making contact with dangerous moving parts. A good safeguarding system

eliminates the possibility of operators or workers placing their hands near hazardous moving parts.

- *Secure.* Workers should not be able to easily remove or tamper with the safeguard, because a safeguard that can easily be made ineffective is no safeguard at all. Guards and safety devices should be made of durable material that will withstand the conditions of normal use. They must be firmly secured to the machine.

Machines often produce noise (unwanted sound), and this can result in a number of hazards to workers. Not only can it startle and disrupt concentration, but it can interfere with communications, thus hindering the worker's safe job performance. Research has linked noise to a whole range of harmful health effects, from hearing loss and aural pain to nausea, fatigue, reduced muscle control, and emotional disturbances. Engineering controls such as the use of sound-dampening materials, as well as less sophisticated hearing protection, such as ear plugs and muffs, have been suggested as ways of controlling the harmful effects of noise. Vibration, a related hazard that can cause noise and thus result in fatigue and illness for the worker, may be avoided if machines are properly aligned, supported, and, if necessary, anchored.

Because some machines require the use of cutting fluids, coolants, and other potentially harmful substances, operators, maintenance workers, and others in the vicinity may need protection. These substances can cause ailments ranging from dermatitis to serious illnesses and disease. Specially constructed safeguards, ventilation, and protective equipment and clothing are possible temporary solutions to the problem of machinery-related chemical hazards until these hazards can be better controlled or eliminated from the workplace. Some safeguards are:

- *Protect from Falling Objects.* The safeguard should ensure that no objects can fall into moving parts. A small tool that is dropped into a cycling machine could easily become a projectile that could strike and injure someone.
- *Create No New Hazards.* A safeguard defeats its own purpose if it creates a hazard of its own such as a shear point, a jagged edge, or an unfinished surface that can cause a laceration. The edges of guards, for instance, should be rolled or bolted in such a way that eliminates sharp edges.
- *Create No Interference.* Any safeguard that impedes a worker from performing the job quickly and comfortably might soon be overridden or disregarded. Proper safeguarding can actually enhance efficiency, since it can relieve the worker's apprehensions about injury.
- *Allow Safe Lubrication.* If possible, one should be able to lubricate the machine without removing the safeguards. Locating oil reservoirs outside the guard, with a line leading to the lubrication point, will reduce the need for the operator or maintenance worker to enter the hazardous area.

5.3 Danger Sources

All power sources for machinery are potential sources of danger. When using electrically powered or controlled machines, for instance, the equipment as well as the electrical system itself must be properly grounded. Replacing frayed, exposed, or old wiring will also help to protect the operator and others from electrical shocks or electrocution. High-pressure systems, too, need careful inspection and maintenance to prevent possible failure from pulsation, vibration, or leaks. Such a failure could cause explosions or flying objects.

6 MACHINE SAFEGUARDING METHODS*

6.1 General Classifications

There are many ways to safeguard machinery. The type of operation, the size or shape of stock, the method of handling the physical layout of the work area, the type of material, and production requirements or limitations all influence selection of the appropriate safeguarding method(s) for the individual machine.

As a general rule, power transmission apparatus is best protected by fixed guards that enclose the danger area. For hazards at the point of operation, where moving parts actually perform work on stock, several kinds of safeguarding are possible. One must always choose the most effective and practical means available.

1. Guards
 - a. Fixed
 - b. Interlocked
 - c. Adjustable
 - d. Self-adjusting
2. Devices
 - a. Presence sensing
 - Photoelectrical (optical)
 - Radio frequency (capacitance)
 - Electromechanical
 - b. Pullback
 - c. Restraint
 - d. Safety controls
 - Safety trip controls
 - Pressure-sensitive body bar
 - Safety trip rod
 - Safety tripwire cable
 - Two-hand control
 - Two-hand trip
 - e. Gates interlocked other
3. Location/distance
4. Potential feeding and ejection methods to improve safety for the operator
 - a. Automatic feed
 - b. Semiautomatic feed
 - c. Automatic ejection
 - d. Semiautomatic ejection
 - e. Robot
5. Miscellaneous aids
 - a. Awareness barriers
 - b. Miscellaneous protective shields
 - c. Hand-feeding tools and holding fixtures

* From J. B. ReVelle, *Engineering Controls: A Comprehensive Overview*. Used by permission of The Merritt Company, Publisher, from T. S. Ferry, *Safety Management Planning*. Copyright © 1982, The Merritt Company, Santa Monica, CA.

Table 5 Machine Safeguarding: Guards

Method	Safeguarding Action	Advantages	Limitations
Fixed	Provides a barrier	Can be constructed to suit many specific applications In-plant construction is often possible Can provide maximum protection Usually requires minimum maintenance Can be suitable to high production, repetitive operations	May interfere with visibility Can be limited to specific operations Machine adjustment and repair often requires its removal, thereby necessitating other means of protection for maintenance personnel
Interlocked	Shuts off or disengages power and prevents starting of machine when guard is open; should require the machine to be stopped before the worker can reach into the danger area	Can provide maximum protection Allows access to machine for removing jams without time-consuming removal of fixed guards	Requires careful adjustment and maintenance May be easy to disengage
Adjustable	Provides a barrier that may be adjusted to facilitate a variety of production operations	Can be constructed to suit many specific applications Can be adjusted to admit varying sizes of stock	Hands may enter danger area; protection may not be complete at all times May require frequent maintenance and/or adjustment The guard may be made ineffective by the operator May interfere with visibility
Self-adjusting	Provides a barrier that moves according to the size of the stock entering danger area	Off-the-shelf guards are often commercially available	Does not always provide maximum protection May interfere with visibility May require frequent maintenance and adjustment

6.2 Guards, Devices, and Feeding and Ejection Methods

Tables 5–7 provide the interested reader with specifics regarding machine safeguarding.

7 ALTERNATIVES TO ENGINEERING CONTROLS*

Engineering controls are an alternative to PPE, or is it the other way around? This chicken-and-egg situation has become an emotionally charged issue with exponents on both sides arguing their beliefs with little in the way of well-founded evidence to support their cases. The reason for this unfortunate situation is that there is no single solution to all the hazardous operations found in industry. The only realistic answer to the question is that it depends. Each and every situation requires an independent analysis considering all the known factors so that a truly unbiased decision can be reached.

This section presents material useful to engineers in the selection and application of solutions to industrial safety and health problems. Safety and health engineering control principles are deceptively few: substitution, isolation, and ventilation, both general and localized. In a technological sense, an appropriate combination of these strategic principles can be brought

* From J. B. ReVelle, *Engineering Controls: A Comprehensive Overview*. Used by permission of The Merritt Company, Publisher, from T. S. Ferry, *Safety Management Planning*. Copyright © 1982, The Merritt Company, Santa Monica, CA.

Table 6 Machine Safeguarding: Devices

Method	Safeguarding Action	Advantages	Limitations
Photoelectric	Machine will not start cycling when the light field is interrupted When the light field is broken by any part of the operator's body during the cycling process, immediate machine braking is activated	Can allow freer movement for operator	Does not protect against mechanical failure May require frequent alignment and calibration Excess vibration may cause lamp filament damage and premature burnout Limited to machines that can be stopped
Radio frequency (capacitance)	Machine cycling will not start when the capacitance field is interrupted When the capacitance field is disturbed by any part of the operator's body during the cycling process, immediate machine braking is activated	Can allow freer movement for operator	Does not protect against mechanical failure Antenna sensitivity must be properly adjusted Limited to machines that can be stopped
Electromechanical	Contact bar or probe travels a predetermined distance between the operator and the danger area Interruption of this movement prevents the starting of machine cycle	Can allow access at the point of operation	Contact bar or probe must be properly adjusted for each application; this adjustment must be maintained properly
Pullback	As the machine begins to cycle, the operator's hands are pulled out of the danger area	Eliminates the need for auxiliary barriers or other interference at the danger area	Limits movement of operator May obstruct workspace around operator Adjustments must be made for specific operators and for each individual Requires frequent inspections and regular maintenance Requires close supervision of the operator's use of the equipment
Restraint (holdback)	Prevents the operator from reaching into the danger area	Little risk of mechanical failure	Limits movements of operator May obstruct workspace Adjustments must be made for specific operations and each individual Requires close supervision of the operator's use of the equipment
Safety trip controls Pressure-sensitive body bar Safety tripod Safety tripwire cable Two-hand control	Stops machine when tripped Concurrent use of both hands is required preventing the operator from entering the danger area	Simplicity of use Operator's hands are at a predetermined location Operator's hands are free to pick up a new part after first half of cycle is completed	All controls must be manually activated May be difficult to activate controls because of their location Only protects the operator May require special fixtures to hold work May require a machine brake Requires a partial cycle machine with a brake Some two-hand controls can be rendered unsafe by holding with arm or blocking, thereby permitting one-hand operation Protects only the operator

Table 6 (Continued)

Method	Safeguarding Action	Advantages	Limitations
Two-hand trip	Concurrent use of two hands on separate controls prevents hands from being in danger area when machine cycle starts	Operator's hands are away from danger area Can be adapted to multiple operations No obstruction to hand feeding Does not require adjustment for each operation	Operator may try to reach into danger area after tripping machine Some trips can be rendered unsafe by holding with arm or blocking, thereby permitting one-hand operation Protects only the operator May require special fixtures
Gates Interlocked Other	Provides a barrier between danger area and operator or other personnel	Can prevent reaching into or walking into the danger area	May require frequent inspection and regular maintenance May interfere with operator's ability to see the work

Table 7 Machine Safeguarding: Feeding and Ejection Methods

Method	Safeguarding Action	Advantages	Limitations
Automatic feed Semiautomatic feed	Stock is fed from rolls, indexed by machine mechanism, etc. Stock is fed by chutes, movable dies, dial feed, plungers, or sliding bolster	Eliminates the need for operator involvement in the danger area	Other guards are also required for operator protection, usually fixed barrier guards Requires frequent maintenance May not be adaptable to stock variation May create a hazard of blowing chips or debris
Automatic ejection	Workpieces are ejected by air or mechanical means		Size of stock limits the use of this method Air ejection may present a noise hazard
Semiautomatic ejection	Workpieces are ejected by mechanical means which are initiated by the operator	Operator does not have to enter danger area to remove finished work	Other guards are required for operator protection
Robots	Perform work usually done by operator	Operator does not have to enter danger area Are suitable for operations where high stress factors are present, such as heat and noise	May not be adaptable to stock variation Can create hazards themselves Require maximum maintenance Are suitable only to specific operations

to bear on any industrial safety or hygiene control problem to achieve a satisfactory quality of the work environment. It usually is not necessary or appropriate to apply all these principles to any specific potential hazard. A thorough analysis of the control problem must be made to ensure that a proper choice from among these methods will produce the proper control in a manner that is most compatible with the technical process, is acceptable to the workers in terms of day-to-day operation, and can be accomplished with optimal balance of installation and operating expenses.

7.1 Substitution

Although frequently one of the most simple engineering principles to apply, substitution is often overlooked as an appropriate solution to occupational safety and health problems. There is a

tendency to analyze a particular problem from the standpoint of correcting rather than eliminating it. For example, the first inclination in considering a vapor exposure problem in a degreasing operation is to provide ventilation of the operation rather than consider substituting a solvent having a much lower degree of hazard associated with its use. Substitution of less hazardous substances, changing from one type of process equipment to another, or, in some cases, even changing the process itself may provide an effective control of a hazard at minimal expense.

This strategy is often used in conjunction with safety equipment: substituting safety glass for regular glass in some enclosures, replacing unguarded equipment with properly guarded machines, replacing safety gloves or aprons with garments made of a material that is more impervious to the chemicals being handled. Since substitution of equipment frequently is done as an immediate response to an obvious problem, it is not always recognized as an engineering control, even though the end result is every bit as effective.

Substituting one process or operation for another may not be considered except in major modifications. In general, a change in any process from a batch to a continuous type of operation carries with it an inherent reduction in potential hazard. This is true primarily because the frequency and duration of potential contact of workers with the process materials are reduced when the overall process approach becomes one of continuous operation. The substitution of processes can be applied on a fundamental basis, for example, substitution of airless spray for conventional spray equipment can reduce the exposure of a painter to solvent vapors. Substitution of a paint dipping operation for the paint spray operation can reduce the potential hazard even further. In any of these cases, the automation of the process can further reduce the potential hazard (see Table 8).

7.2 Isolation

Application of the principle of isolation is frequently envisioned as consisting of the installation of a physical barrier (such as a machine guard or device—refer to Tables 5 and 6) between a hazardous operation and the workers. Fundamentally, however, isolation can be provided *without* a physical barrier through the appropriate use of distance and, in some situations, time.

Table 8 Positive Performance Characteristics—Some Things Done Better by People versus Machines

People	Machines
Detect signals in high-noise fields	Respond quickly to signals
Recognize objects under widely different conditions	Sense energies outside human range
Perceive patterns	Consistently perform precise, routine, repetitive operations
Sensitive to a wide variety of stimuli	Recall and process enormous amounts of data
Long-term memory	Monitor people or other machines
Handle unexpected or low-probability events	Reason deductively
Reason inductively	Exert enormous power
Profit from experience	Relatively uniform performance
Exercise judgment	Rapid transmission of signals
Flexibility, improvisation, and creativity	Perform several tasks simultaneously
Select and perform under overload conditions	Expendable
Adapt to changing environment	Resistance to many environmental stresses
Appreciate and create beauty	
Perform fine manipulations	
Perform when partially impaired	
Relatively maintenance free	

Perhaps the most common example of isolation as a control strategy is associated with storage and use of flammable solvents. The large tank farms with dikes around the tanks, underground storage of some solvents, the detached solvent sheds, and fireproof solvent storage rooms within buildings are all commonplace in American industry. Frequently, the application of the principle of isolation maximizes the benefits of additional engineering concepts such as excessive noise control, remote control materials handling (as with radioactive substances), and local exhaust ventilation.

7.3 Ventilation

Workplace air quality is affected directly by the design and performance of the exhaust system. An improperly designed hood or a hood evacuated with an insufficient volumetric rate of air will contaminate the occupational environment and affect workers in the vicinity of the hazard source. This is a simple, but powerful, symbolic representation of one form of the close relationship between atmospheric emissions (as regulated by the EPA) and occupational exposure (as regulated by OSHA). What is done with gases generated as a result of industrial operations/processes? These emissions can be exhausted directly to the atmosphere, indirectly to the atmosphere (from the workplace through the general ventilation system), or recirculated to the workplace. The effectiveness of the ventilation system design and operation impacts directly on the necessity and type of respiratory gear needed to protect the workforce.

8 DESIGN AND REDESIGN*

8.1 Hardware

Designers of machines must consider the performance characteristics of machine operators as a major constraint in the creation or modification of both mechanical and electrical equipment. To do less would be tantamount to ignoring the limitations of human capabilities.

Equipment designers especially concerned with engineering controls to be incorporated into machines, whether at the time of initial conceptualization or later when alterations are to be made, must also be cognizant of the principles of human factors (ergonomics). Equipment designers are aware that there are selected tasks that people can perform with greater skill and dependability than machines, and vice versa. Some of these positive performance characteristics are noted in Table 8. In addition, designers of equipment and engineering controls are knowledgeable of human performance limitations, both physically and psychologically. They know that the interaction of forces between people and their operating environment presents a never-ending challenge in assessing the complex interrelationships that provide the basis for that often fine line between safety versus hazard or health versus contaminant. Table 9 identifies the six pertinent sciences most closely involved in the design of machines and engineering controls. It is both rational and reasonable to expect that, when engineering controls are being considered to eliminate or reduce hazards or contaminants, designers make full use of the principles established by specialists in these human performance sciences.

8.2 Process

A stress (or stressor) is some physical or psychological feature of the environment that requires an operator to be unduly exerted to continue performing. Such exertion is termed strain as in

* From J. B. ReVelle, *Engineering Controls: A Comprehensive Overview*. Used by permission of The Merritt Company, Publisher, from T. S. Ferry. *Safety Management Planning*. Copyright © 1982, The Merritt Company, Santa Monica, CA.

Table 9 People Performance Sciences

Anthropometry	Pertains to the measurement of physical features and characteristics of the static human body
Biomechanics	A study of the range, strength, endurance, and accuracy of movements of the human body
Ergonomics	Human factors engineering, especially biomechanics aspects
Human factors engineering	Designing for human use
Kinesiology	A study of the principles of mechanics and anatomy of human movement
Systems safety engineering	Designing that considers the operator's qualities, the equipment, and the environment relative to successful task performance

“stress and strain.” Common physical stressors in industrial workplaces are poor illumination, excessive noise, vibration, heat, and the presence of excessive, harmful atmospheric contaminants. Unfortunately, much less is known about their effects when they occur at the same time, in rapid sequence, or over extended periods of time. Research suggests that such effects are not simply additive, but synergistic, thus compounding their detrimental effects. In addition, when physical work environments are unfavorable to equipment operators, two or more stressors are generally present: high temperature and excessive noise, for example. The solution to process design and redesign is relatively easy to specify but costly to implement—design the physical environment so that all physical characteristics are within an acceptable range.

Marketed in the United States since the early 1960s, industrial robots offer both hardware and process designers a technology that can be used when hazardous or uncomfortable working conditions are expected or already exist. Where a job situation poses potential dangers or the workplace is hot or in some other way unpleasant, a robot should be considered as a substitute for human operators. Hot forging, die casting, and spray painting fall into this category. If work parts or tools are awkward or heavy, an industrial robot may fill the job. Some robots are capable of lifting items weighing several hundred pounds.

An industrial robot is a general-purpose, programmable machine that possesses certain humanlike capabilities. The most obvious characteristic is the robot's arm, which, when combined with the robot's capacity to be programmed, makes it ideally suited to a variety of uncomfortable/undesirable production tasks. Hardware and process designers now possess an additional capability for potential inclusion in their future designs and redesigns.

8.3 Hazardous Material Classification System

The NFPA, a private, nonprofit organization, is the leading authoritative source of technical background, data, and consumer advice on fire protection, problems, and prevention. The primary goal of the NFPA is to reduce the worldwide burden of fire and other hazards on the quality of life by providing and advocating scientifically based consensus codes and standards, research, training, and education. The NFPA has in excess of 300 codes worldwide, which are for sale through their website, <http://www.nfpa.org/>. While NFPA codes cover several aspects of flammable and otherwise hazardous materials, perhaps the most significant is the *NFP 704 Hazard Identification* ratings system (the familiar NFPA “hazard diamond” for health, flammability, and instability (see Fig. 2).

What do the numbers and symbols on an NFPA fire diamond mean? The diamond is divided into four sections. Numbers in the three colored sections range from 0 (least severe hazard) to 4 (most severe hazard). The fourth (white) section is left blank and is used only to denote special firefighting measures/hazards.

Health hazard levels are identified in the upper left section (blue in the original) of the diamond (third base) as follows:

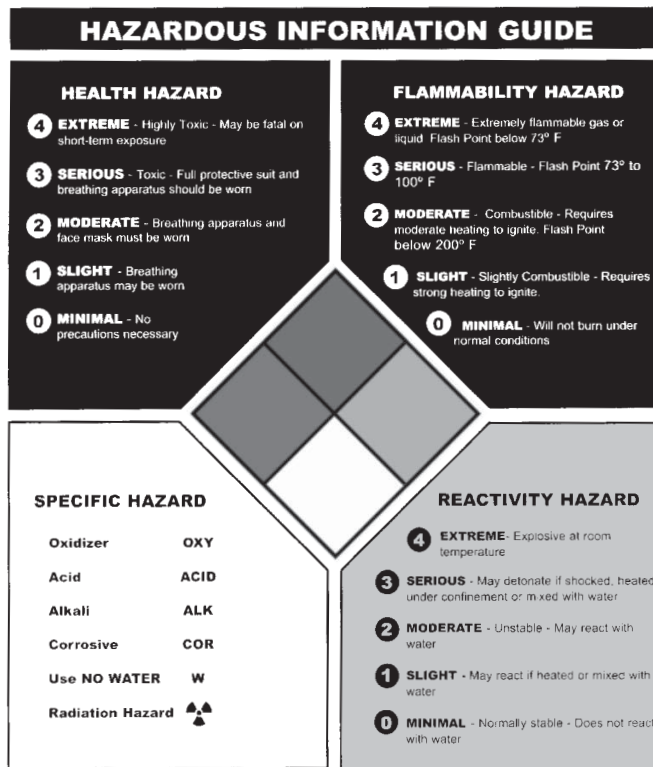


Figure 2 NFPA hazard diamond. In the color original, the upper left section is blue, the upper right is red, the bottom right is yellow, and the bottom left is white.

- *Level 4* is used when even very short exposure could cause death or serious residual injury even though prompt medical attention is given.
- *Level 3* is used when short exposure could cause serious temporary or residual injury even though prompt medical attention is given.
- *Level 2* is used when intense or continued exposure could cause temporary incapacitation or possible residual injury unless prompt medical attention is given.
- *Level 1* is used when exposure could cause irritation but only minor residual injury even if no treatment is given.
- *Level 0* is used when exposure under fire conditions would offer no hazard beyond that of ordinary combustible materials.

Flammability hazard levels are identified in the upper right section (red in the original) of the diamond (second base) as follows:

- *Level 4* is used when a material will rapidly or completely vaporize at normal pressure and temperature or is readily dispersed into the air and will burn readily.
- *Level 3* is used when a material is a solid or liquid that can be ignited under almost all ambient conditions.
- *Level 2* is used when a material must be moderately heated or exposed to relatively high temperature before ignition can occur.

- *Level 1* is used when a material must be preheated before ignition can occur.
- *Level 0* is used when a material will not burn.

*Instability** hazard levels are identified in the bottom right section (yellow in the original) of the diamond (first base) as follows:

- *Level 4* is used when a material is readily capable of detonation or of explosive decomposition or reaction at normal temperatures and pressures.
- *Level 3* is used when a material is capable of detonation or explosive reaction but requires a strong initiating source or must be heated under confinement before initiation or reacts explosively with water.
- *Level 2* is used when a material is normally unstable and readily undergoes violent decomposition but does not detonate. Also may react violently with water or may form potentially explosive mixtures with water.
- *Level 1* is used when a material is normally stable but can become unstable at elevated temperatures and pressures or may react with water with some release of energy but not violently.
- *Level 0* is used when a material is normally stable, even under fire exposure conditions, and is not reactive with water.

Special† hazard levels are identified in the bottom left section (white in the original) of the diamond (home plate) as follows:

- *Symbol OX* is used when a material is an oxidizer, a chemical that can greatly increase the rate of combustion/fire.
- *Symbol W* is used when a material has unusual reactivity with water. This indicates a potential hazard when using water to fight a fire involving this material.
- *ACID* is used when a material is an acid, i.e., a corrosive substance that has a pH (power of hydrogen) lower than 7.0.
- *ALK* is used when a material is an alkaline substance, also referred to as a base. These caustic materials have a pH greater than 7.0.
- *COR* is used when a material is corrosive. It could be either an acid or a base.
- *XXX* is used as another symbol for corrosive.
- *SAC* is used when a material is a poison or highly toxic substance.
- *RAD* is used when a material is a radioactive hazard. Radioactive materials are extremely hazardous when inhaled or ingested.
- *EXP* is used when a material is explosive. This symbol is redundant since explosives are easily recognized by their *instability* rating.

Readers who desire to obtain an NFPA hazard diamond wall chart (see Fig. 2) at no cost can do so by contacting Graphic Products, Inc. on the Internet at <http://www.graphicproducts.com>.

* Prior to 1996, this section was titled *Reactivity*. The name was changed because many people did not understand the distinction between a “reactive hazard” and the “chemical reactivity” of a material. The numeric ratings and their meanings remain unchanged.

† There are only two *NFPA 704* approved symbols, *OX* and *W*. Other symbols, abbreviations, and words that some organizations used in the white *Special* hazards section are also described. These uses are *not compliant* with *NFPA 704* but are presented here in case they appear on a material safety data sheet, which is discussed in Section 8.4, or a container label.

8.4 Material Safety Data Sheets

Material safety data sheets (MSDSs) are designed to provide both workers and emergency personnel with the proper procedures for handling or working with a particular substance. MSDSs include information such as physical data (melting point, boiling point, flash point, etc.), toxicity, health effects, first-aid, reactivity, storage, disposal, protective equipment, and spill/leak procedures. These are of particular use if a spill or other accident occurs.

MSDSs Are Meant For

- Employees who may be occupationally exposed to a hazard at work
- Employees who need to know the proper methods for storage
- Emergency responders, such as fire fighters, hazardous material crews, emergency medical technicians, and emergency room personnel

MSDSs Are Not Meant For

- *Consumers.* An MSDS reflects the hazards of working with specific substances in an occupational situation, *not* by a retail consumer. For example, an MSDS for a can of paint is not pertinent to someone who uses it once a year but is extremely important to someone who is exposed to paint in a confined space 40 h per week.

Sources of MSDSs

- Your laboratory or workplace should have a collection of MSDSs that came with the hazardous chemicals you have ordered. *Note:* Do not throw them away; file them where they can be easily located in an emergency.
- Most universities and businesses have a collection of MSDSs somewhere on-site. Check with your Environmental or Occupational Safety and Health Office or your science librarian. Some organizations use commercial services to obtain printed, FAX, or on-line copies of MSDSs.
- MSDSs can be obtained from the distributor that sold you the material. If they cannot be located, try the manufacturer's customer service department.
- The Internet has a wide range of free sources. A handy list of 100 such sites can be found at <http://www.ILPL.com/MSDS/Index.html#Internet>.
- MSDS software or Internet subscription services can be purchased.

MSDS Background

- OSHA began requiring MSDSs for hazardous materials on May 26, 1986.
- OSHA is responsible for the Hazard Communication Standard (HCS) 29 CFR (*Code of Federal Regulations*) 1910.1200. The purpose of this standard is "to ensure that the hazards of all chemicals produced or imported are evaluated and that information concerning their hazards is transmitted to employers and employees. This transformation of information is to be accomplished by means of comprehensive hazard communication programs, which are to include container labeling and other forms of warning, material safety data sheets, and employee training." The HCS specifies the required elements that must be on an MSDS, among other important data.

8.5 Safety Design Requirements

In the course of ensuring that product specifications are met, engineers/designers are typically responsible for adhering to a variety of some quite general and some very specific safety design requirements. Since its origins in 1970, OSHA has adopted or originated

numerous safety and health-related requirements. If the reader were to enter the Internet at the website http://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=FEDERAL_REGISTER&pid= and then search for “safety design requirements,” nearly 1300 individual documents would be presented for investigation. Each of these documents can be double-clicked to obtain more detailed information regarding the specific subject matter.

9 PERSONAL PROTECTIVE EQUIPMENT*

9.1 Background

Engineering controls, which eliminate the hazard at the source and do not rely on the worker’s behavior for their effectiveness, offer the best and most reliable means of safe-guarding. Therefore, engineering controls must be the first choice for eliminating machinery hazards. But whenever an extra measure of protection is necessary, operators must wear PPE. If it is to provide adequate protection, the protective clothing and equipment selected must always be:

- Appropriate for the particular hazards
- Maintained in good condition
- Properly stored when not in use to prevent damage or loss
- Kept clean and sanitary

Protective clothing is available for every part of the human body. Hard hats can protect the head from falling objects when the worker is handling stock; caps and hair nets can help keep the worker’s hair from being caught in machinery. If machine coolants could splash or particles could fly into the operator’s eyes or face, then face shields, safety goggles, glasses, or similar kinds of protection must be used. Hearing protection may be needed when workers operate noisy machinery. To guard the trunk of the body from cuts or impacts from heavy or rough-edged stock, there are certain protective coveralls, jackets, vests, aprons, and full-body suits. Workers can protect their hands and arms from the same kinds of injury with special sleeves and gloves. And safety shoes and boots or other acceptable foot guards can shield the feet against injury in case the worker needs to handle heavy stock that might drop.

Protective clothing and equipment themselves can create hazards. Protective glove that can become caught between rotating parts or respirator face pieces that hinder the wearer’s vision require alertness and careful supervision whenever they are used. Other aspects of the worker’s dress may present additional safety hazards. Loose-fitting clothing might become entangled in rotating spindles or other kinds of moving machinery. Jewelry, such as bracelets and rings, can catch on machine parts or stock and lead to serious injury by pulling a hand into the danger area.

Naturally, each situation will vary. In some simple cases, respirators, chemical goggles, aprons, and gloves may be sufficient PPE to afford the necessary coverage. In more complicated situations, even the most sophisticated equipment may not be enough and engineering controls would become mandatory. Safety, industrial, and plant engineers should be expected to provide the necessary analyses to ascertain the extent of the hazard to employees whose work causes them to be exposed to the corrosive fumes.

* From J. B. ReVelle, *Engineering Controls: A Comprehensive Overview*. Used by permission of The Merritt Company, Publisher, from T. S. Ferry, *Safety Management Planning*. Copyright © 1982, The Merritt Company, Santa Monica, CA.

9.2 Planning and Implementing the Use of Protective Equipment*

This section reviews ways to help plan, implement, and maintain PPE. This can be considered in terms of the following nine phases: (1) need analysis, (2) equipment selection, (3) program communication, (4) training, (5) fitting and adjustment, (6) target date setting, (7) break-in period, (8) enforcement, and (9) follow-through.

The first phase of promoting the use of PPE is called *need analysis*. Before selecting protective equipment, the hazards or conditions the equipment must protect the employee from must be determined. To accomplish this, questions such as the following must be asked:

- What standards does the law require for this type of work in this type of environment?
- What needs do our accident statistics point to?
- What hazards have we found in our safety and/or health inspections?
- What needs show up in our job analysis and job observation activities?
- Where is the potential for accidents, injuries, illnesses, and damage?
- Which hazards cannot be eliminated or segregated?

The second phase of promoting the use of protective equipment is *equipment selection*. Once a need has been established, proper equipment must be selected. Basic consideration should include the following:

- Conformity to the standards
- Degree of protection provided
- Relative cost
- Ease of use and maintenance
- Relative comfort

The third phase is *program communication*. It is not appropriate to simply announce a protective equipment program, put it into effect, and expect to get immediate cooperation. Employees tend to resist change unless they see it as necessary, comfortable, or reasonable. It is helpful to use various approaches to publicity and promotion to teach employees why the equipment is necessary. Various points can be covered in supervisor's meetings, in safety meetings, by posters, on bulletin boards, in special meetings, and in casual conversation. Gradually, employees will come to expect or to request protective equipment to be used on the job. The main points in program communication are to educate employees in why protective equipment is necessary and to encourage them to want it and to use it.

Training is an essential step in making sure protective equipment will be used properly. The employees should learn why the equipment is necessary, when it must be used, who must use it, where it is required, what the benefits are, and how to use it and take care of it. Do not forget that employee turnover will bring new employees into the work area. Therefore, you will continually need to train new employees in the use of the protective equipment they will handle.

After the training phase comes the *fitting and adjustment* phase. Unless the protective equipment fits the individual properly, it may not give the necessary protection. There are many ways to fit or to adjust protective equipment. For example, face masks have straps that hold

* From J. B. ReVelle and J. Stephenson, *Safety Training Methods*, 2nd ed. Copyright © 1995. Reprinted by permission of John Wiley & Sons, New York.

them snug against the contours of the head and face and prevent leaks; rubberized garments have snaps or ties that can be drawn up snugly, to keep loose and floppy garments from getting caught in machinery.

The next phase is *target date setting*. After the other phases have been completed, set specific dates for completion of the various phases. For example, all employees should be fitted with protective equipment before a certain date; all training should be completed by a certain date; after a certain date, all employees must wear their protective equipment while in the production area.

After setting the target dates, expect a *break-in period*. There will usually be a period of psychological adjustment whenever a new personal protective program is established. Remember two things:

- Expect some gripes, grumbles, and problems.
- Appropriate consideration must be given to each individual problem; then strive toward a workable solution.

It might also be wise to post signs that indicate the type of equipment needed. For example, a sign might read, "Eye protection must be worn in this area."

After the break-in phase comes *enforcement*. If all the previous phases were successful, problems in terms of enforcement should be few. In case disciplinary action is required, sound judgment must be used and each case must be evaluated on an individual basis. If employees fail to use protective equipment, they may be exposed to hazards. Do not forget, the employer can be penalized if employees do not use their protection.

The final phase is *follow-through*. Although disciplinary action may sometimes be necessary, positive motivation plays a more effective part in a successful protective equipment program. One type of positive motivation is a proper example set by management. Managers must wear their protective equipment, just as employees are expected to wear theirs.

9.3 Adequacy, Maintenance, and Sanitation

Before selling safety shoes and supplying safety goggles at a company store, the attendants must be guided by a well-structured program of equipment maintenance, preferably preventive maintenance. Daily maintenance of different types of equipment might include adjustment of the suspension system on a safety hat; cleaning of goggle lenses, glasses, or spectacles; scraping residue from the sole of a safety shoe; or proper adjustment of a face mask when donning an air-purifying respirator.

Performing these functions should be coupled with periodic inspections for weaknesses or defects in the equipment. How often this type of check is made, of course, depends on the particular type of equipment used. For example, sealed-canister gas masks should be weighed on receipt from the manufacturer and the weight should be marked indelibly on each canister. Stored units should then be reweighed periodically, and those exceeding a recommended weight should be discarded even though the seal remains unbroken.

Sanitation, as spelled out in the OSHA Act, is a key part of any operation. It requires the use of PPE, not only to eliminate cross-infection among users of the same unit of equipment, but because unsanitary equipment is objectionable to the wearer. Procedures and facilities that are necessary to sanitize or disinfect equipment can be an integral part of an equipment maintenance program. For example, the OSHA Act says, "Respirators used routinely shall be inspected during cleaning." Without grime and dirt to hinder an inspection, gauges can be read better, rubber or elastomer parts can be checked for pliability and signs of deterioration, and valves can be checked.

10 MANAGING THE SAFETY FUNCTION

10.1 Supervisor's Role

The responsibilities of the first-line supervisor are many. Direction of the workforce includes the following supervisory functions:

- Setting goals
- Improving present work methods
- Delegating work
- Allocating manpower
- Meeting deadlines
- Controlling expenditures
- Following progress of work
- Evaluating employee performance
- Forecasting manpower requirements
- Supervising on-the-job training
- Reviewing employee performance
- Handling employee complaints
- Enforcing rules
- Conducting meetings
- Increasing safety awareness*

Supervisory understanding of the interrelationships of these responsibilities is a learned attribute. Organizations that expect their supervisors to offer a high quality of leadership to their employees must provide appropriate training and experiential opportunities to current supervisors and supervisory trainees alike.

10.2 Elements of Accident Prevention[†]

- Safety policy must be clearly defined and communicated to all employees.
- The safety record of a company is a barometer of its efficiency. An American Engineering Council study revealed that “maximum productivity is ordinarily secured only when the accident rate tends toward the irreducible minimum.”
- Unless line supervisors are accountable for the safety of all employees, no safety program will be effective. Top management must let all supervisors and managers know what is expected of them in safety.
- Periodic progress reports are required to let managers and employees know what they have accomplished in safety.
- Meetings with supervisors and managers to review accident reports, compensation costs, accident–cause analysis, and accident prevention procedures are important elements of the overall safety program.
- The idea of putting on a big safety campaign with posters, slogans, and safety contests is wrong. The Madison Avenue approach does not work over the long run.

* From B. D. Lewis, Jr., *The Supervisor in 1975*. Copyright © September 1973. Reprinted with permission of *Personnel Journal*, Costa Mesa, CA.

† From R. De Reamer, *Modern Safety and Health Technology*. Copyright © 1980. Reprinted by permission of John Wiley & Sons, New York.

- Good housekeeping and the enforcement of safety rules show that management has a real concern for employee welfare. They are important elements in the development of good morale. (A U.S. Department of Labor study has revealed that workers are vitally concerned with safety and health conditions of the workplace. A surprisingly high percentage of workers ranked protection against work-related injuries and illness and pleasant working conditions as having a priority among their basic on-the-job needs. In fact, they rated safety higher than fringe benefits and steady employment.)
- The use of PPE (safety glasses, safety shoes, hard hats, etc.) must be a condition of employment in all sections of the plant where such protection is required.
- Safety files must be complete and up to date to satisfy internal information requirements as well as external inspections by OSHA compliance officers and similar officials (see Table 10).

10.3 Management Principles*

- Regardless of the industry or the process, the role of supervisors and managers in any safety program takes precedence over any of the other elements. This is not to say that the managerial role is necessarily more important than the development of safe environments, but without manager and supervisor participation, the other elements have a lukewarm existence. There is a dynamic relationship between management and the development of safe working conditions and management and the development of safety awareness, and the relationship must not be denied. The items in Table 10 are presented for use during the review of the administrative storage index to determine the adequacy of safety-related files.
- Where responsibility for preventing accidents and providing a healthful work environment is sloughed off to the safety department or a safety committee, any reduction in the accident rate is minimal. To reduce the accident rate and, in particular, to make a good rate better, line managers must be held responsible and accountable for safety. Every member of the management team must have a role in the safety program. Admittedly, this idea is not new, but application of the concept still requires crystal-clear definition and vigorous promotion.
- Notwithstanding the many excellent examples of outstanding safety records that have been achieved because every member of management had assumed full responsibility for safety, there are still large numbers of companies, particularly the small establishments, using safety contests, posters, or safety committees as the focal point of their safety programs—but with disappointing results. Under such circumstances safety is perceived as an isolated aspect of the business operation with rather low ceiling possibilities at best. But there are some who feel that gimmicks must be used because foremen and the managers do not have time for safety.
- As an example of the case in point, a handbook on personnel contains the statement that “A major disadvantage of some company-sponsored safety programs is that the supervisor can’t spare sufficient time from his regular duties for running the safety program.” Significantly, this was not a casual comment in a chapter on safety. It was indented and in bold print for emphasis. Yet it is a firmly accepted fact that to achieve good results in safety, managers and supervisors must take the time to fulfill their safety responsibilities. Safety is one of their regular duties.
- The interrelationships of the many components of an effective industrial safety program are portrayed in Fig. 3.

* From R. De Reamer, *Modern Safety and Health Technology*. Copyright © 1980. Reprinted by permission of John Wiley & Sons, New York.

Table 10 Requirements for Safety Files

Number	Action Required	Action Completed	
		Yes	No
1.	Is there a separate section for safety-related files?		
2.	Are the following subjects provided for in the safety section of the files:		
	a. Blank OSHA forms?		
	b. Completed OSHA forms?		
	c. Blank company safety forms?		
	d. Completed company safety forms?		
	e. Blank safety checklists?		
	f. Completed safety checklists?		
	g. Agendas of company safety meetings?		
	h. Minutes of company safety meetings?		
	i. Records of safety equipment purchases?		
	j. Records of safety equipment checkouts?		
	k. Incoming correspondence related to safety?		
	l. Outgoing correspondence related to safety?		
	m. Record of safety projects assigned?		
	n. Record of safety projects completed?		
	o. Record of fire drills (if applicable)?		
	p. Record of external assistance used to provide specialized safety expertise?		
	q. Record of inspections by fire department, insurance companies, state and city inspectors, and OSHA compliance officers?		
	r. National Safety Council catalogs and brochures for films, posters, and other safety-related materials?		
3.	Are the files listed in item 2 reviewed periodically:		
	a. To ensure that they are current?		
	b. To retire material over five years old?		
4.	Are safety-related files reviewed periodically to determine the need to eliminate selected files and to add new subjects?		
5.	Is the index to the file current, so that an outsider could easily understand the system?		

Source: J. B. ReVelle and J. Stephenson, *Safety Training Methods*, 2nd ed. Copyright © 1995. Reprinted by permission of John Wiley & Sons, New York.

10.4 Eliminating Unsafe Conditions*

The following steps should be taken to effectively and efficiently eliminate an unsafe condition:

- *Remove.* If at all possible, have the hazard eliminated.
- *Guard.* If the danger point (e.g., high-tension wires) cannot be removed, see to it that hazard is shielded by screens, enclosures, or other guarding devices.
- *Warn.* If guarding is impossible or impractical, warn of the unsafe condition. If a truck must back up across a sidewalk to a loading platform, the sidewalk cannot be removed

* From J. B. ReVelle and J. Stephenson, *Safety Training Methods*, 2nd ed. Copyright © 1995. Reprinted by permission of John Wiley & Sons, New York

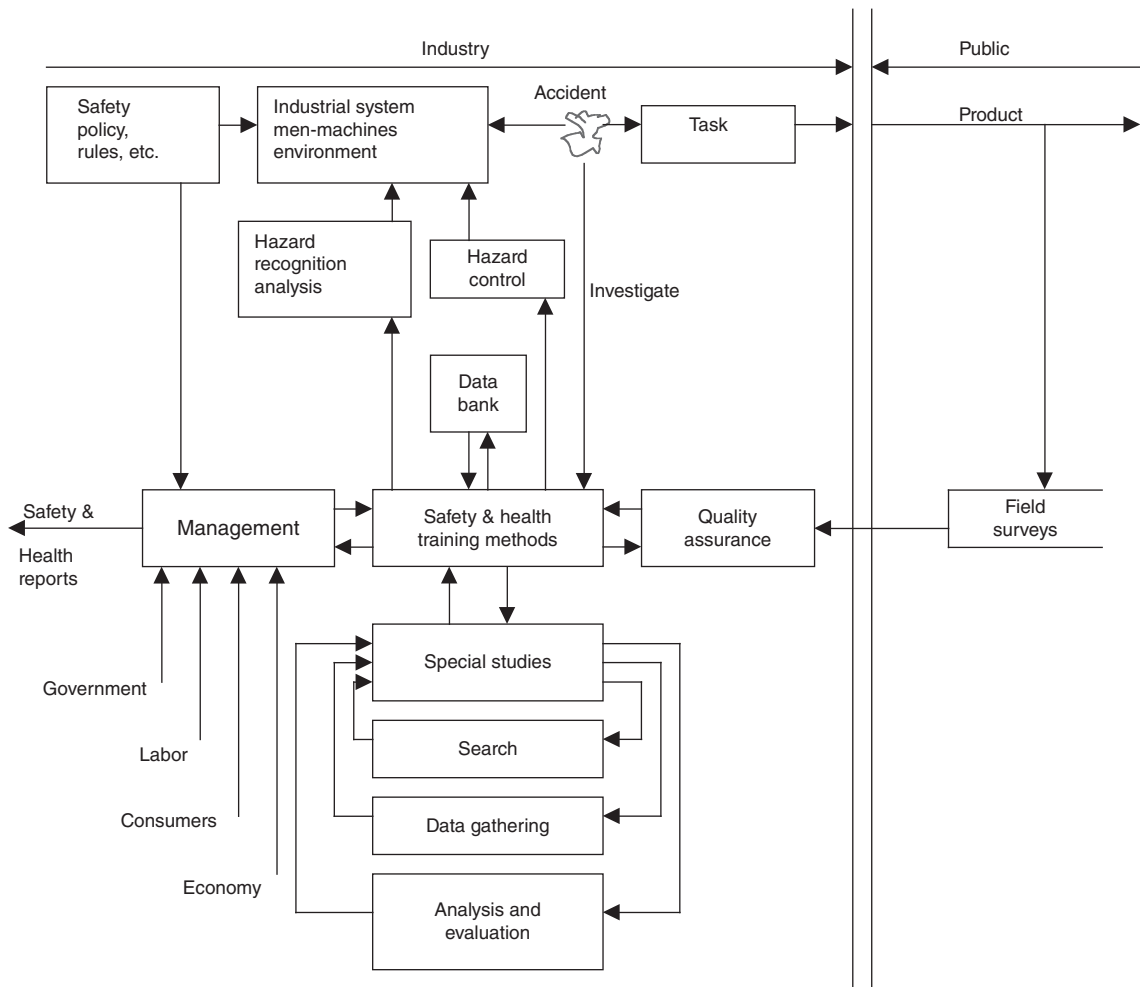


Figure 3 Basic functions of an effective safety program. *Source:* Reprinted with permission from *Industrial Engineering Magazine*. Copyright © 1979 American Institute of Industrial Engineers, Inc., Norcross, GA.

or a fence built around the truck. All that can be done is to warn that an unsafe condition exists. This is done by posting a danger sign or making use of a bell, horn, whistle, signal light, painted striped lines, red flag, or other device.

- *Recommend.* If you cannot remove or guard an unsafe condition on your own, notify the proper authorities about it. Make specific recommendations as to how the unsafe condition can be eliminated.
- *Follow Up.* After a reasonable length of time, check to see whether the recommendation has been acted on or whether the unsafe condition still exists. If it remains, the person or persons to whom the recommendations were made should be notified.

Five S (5S)

The original 5S principles were stated in Japanese. Because of their proven value to western industry, they have been translated and restated in English. The 5S is a mantra of sorts designed to help build a quality work environment, both physical and mental.

The 5S condition of a work area is critical to the morale of employees and is the basis of a customer's first impressions. Management's attitude regarding their employees is reflected in the 5S condition of the work area. The 5S philosophy applies in any work area.

The elements of 5S are simple to learn and important to implement:

- *Sort*. Eliminate whatever is not needed.
- *Straighten*. Organize whatever remains.
- *Shine*. Clean the work area.
- *Standardize*. Schedule regular cleaning and maintenance.
- *Sustain*. Make 5S a way of life.

Numerous benefits can be derived from implementing 5S. Some examples are:

- Improved safety
- Higher equipment availability
- Lower defect rates
- Reduced costs
- Increased production agility and flexibility
- Improved employee morale
- Better asset utilization
- Enhanced enterprise image to customers, suppliers, employees, and management

Table 11 is an example of a 5S workplace scan diagnostic checklist.

The following factors should be considered in organizing a plant that provides for maximum productivity and employee well-being:

- The general arrangement of the facility should be efficient, orderly, and neat.
- Workstations should be clearly identified so that employees can be assigned according to the most effective working arrangement.
- Material flow should be designed to prevent unnecessary employee movement for given work.
- Materials storage, distribution, and handling should be routinized for efficiency and safety.
- Decentralized tool storage should be used wherever possible. Where centralized storage is essential (e.g., general supply areas, locker areas, and project storage areas), care should be given to establish a management system that will avoid unnecessary crowding or congested traffic flow. (Certain procedures, such as time staggering, may reduce congestion.)
- Time use plans should be established for frequently used facilities to avoid having workers wait for a particular apparatus.
- A warning system and communications network should be established for emergencies such as fire, explosion, storm, injuries, and other events that would affect the well-being of employees.

The following unsafe conditions checklist presents a variety of undesirable characteristics to which both employers and employees should be alerted:

- *Unsafe Conditions—Mechanical Failure*. These are types of unsafe conditions that can lead to occupational accidents and injuries. *Note*: Keep in mind that unsafe conditions often come about as a result of unsafe acts.
- *Lack of Guards*. This applies to hazardous places like platforms, catwalks, or scaffolds where no guardrails are provided; power lines or explosive materials that are not fenced

Table 11 5S Diagnostic Checklist

5S Category	5S Item	5S Rating Level					Remarks												
		L0	L1	L2	L3	L4													
Sort (organization)	Distinguish between what is needed and not needed	<table border="1"> <thead> <tr> <th>Number of Problems</th> <th>Rating Level</th> </tr> </thead> <tbody> <tr> <td>3 or more</td> <td>Level 0 (L0)</td> </tr> <tr> <td>3-4</td> <td>Level 1 (L1)</td> </tr> <tr> <td>2</td> <td>Level 2 (L2)</td> </tr> <tr> <td>1</td> <td>Level 3 (L3)</td> </tr> <tr> <td>None</td> <td>Level 4 (L4)</td> </tr> </tbody> </table>	Number of Problems	Rating Level	3 or more	Level 0 (L0)	3-4	Level 1 (L1)	2	Level 2 (L2)	1	Level 3 (L3)	None	Level 4 (L4)					
	Number of Problems		Rating Level																
	3 or more		Level 0 (L0)																
	3-4		Level 1 (L1)																
	2		Level 2 (L2)																
1	Level 3 (L3)																		
None	Level 4 (L4)																		
Unneeded equipment, tools, furniture, etc., are present																			
Unneeded items are on walls, bulletin boards, etc.																			
Items are present in aisles, stairways, corners, etc.																			
Unneeded inventory, supplies, arts, or materials are present																			
Set in Order (orderliness)	Safety hazards (water, oil, chemical, machines) exist																		
	A place for everything and everything in its place																		
	Correct places for items are not obvious																		
	Items are not in their places																		
	Aisles, workstations, equipment locations are not indicated																		
Shine (cleanliness)	Items are not put away immediately after use																		
	Height and quantity limits are not obvious																		
	Cleaning and looking for ways to keep it clean and organized																		
	Floors, walls, stairs, and surfaces are not free of dirt, oil, and grease																		
	Equipment is not kept clean and free of dirt, oil, and grease																		
Standardize (adherence)	Cleaning materials are not easily accessible																		
	Lines, labels, signs, etc., are not clean and unbroken																		
	Other cleaning problems of any kind are present																		
	Maintain and monitor the first three categories																		
	Necessary information is not visible																		
Sustain (self-discipline)	All standards are not known and visible																		
	Checklists do not exist for cleaning and maintenance jobs																		
	All quantities and limits are not easily recognizable																		
	How many items cannot be located in 30 seconds?																		
	Stick to the rules																		
TOTAL																			

off or enclosed in some way; and machines or other equipment where moving parts or other danger points are not safeguarded.

- *Inadequate Guards.* Often a hazard that is partially guarded is more dangerous than it would be if there were no guards. The employee, seeing some sort of guard, may feel secure and fail to take precautions that would ordinarily be taken if there were no guards at all.
- *Defects.* Equipment or materials that are worn, torn, cracked, broken, rusty, bent, sharp, or splintered; buildings, machines, or tools that have been condemned or are in disrepair.
- *Hazardous Arrangement (Housekeeping).* Cluttered floors and work areas: improper layout of machines and other production facilities; blocked aisle space or fire exits; unsafely stored or piled tools and material; overloaded platforms and vehicles; inadequate drainage and disposal facilities for waste products. The reader is referred to the earlier discussion about 5S.
- *Improper Illumination.* Insufficient light; too much light; lights of the wrong color; glare; arrangement of lighting systems that result in shadows and too much contrast.

- *Unsafe Ventilation.* Concentration of vapors, dusts, gases, fumes; unsuitable capacity, location, or arrangement of ventilation system; insufficient air changes, impure air source used for air changes; abnormal temperature and humidity.

In describing conditions for each item to be inspected, terms such as the following should be used:

Broken	Leaking
Corroded	Loose (or slipping)
Decomposed	Missing
Frayed	Rusted
Fuming	Spillage
Gaseous	Vibrating
Jagged	

An alphabetized listing of possible problems to be inspected is presented in Table 12.

Hazard Classification

It is important to differentiate the *degrees of severity* of different hazards. The commonly used standards are given below:

Table 12 Possible Problems to Be Inspected

Acids	Dusts	Railroad cars
Aisles	Electric motors	Ramps
Alarms	Elevators	Raw materials
Atmosphere	Explosives	Respirators
Automobiles	Extinguishers	Roads
Barrels	Flammables	Roofs
Bins	Floors	Safety devices
Blinker lights	Forklifts	Safety glasses
Boilers	Fumes	Safety shoes
Borers	Gas cylinders	Scaffolds
Buggies	Gas engines	Shafts
Buildings	Gases	Shapers
Cabinets	Hand tools	Shelves
Cables	Hard hats	Sirens
Carboys	Hoists	Slings
Catwalks	Hoses	Solvents
Caustics	Hydrants	Sprays
Chemicals	Ladders	Sprinkler systems
Claxons	Lathes	Stairs
Closets	Lights	Steam engines
Connectors	Mills	Sumps
Containers	Mists	Switches
Controls	Motorized carts	Tanks
Conveyors	Piping	Trucks
Cranes	Pits	Vats
Crossing lights	Platforms	Walkways
Cutters	Power tools	Walls
Docks	Presses	Warning devices
Doors	Racks	

Source: Principles and Practices of Occupational Safety and Health: A Programmed Instruction Course, OSHA 2213, Student Manual Booklet 1, U.S. Department of Labor, Washington, DC, p. 40.

- *Class A Hazard.* Any condition or practice with *potential* for causing loss of life or body part and/or extensive loss of structure, equipment, or material.
- *Class B Hazard.* Any condition or practice with *potential* for causing serious injury, illness, or property damage, but less severe than class A.
- *Class C Hazard.* Any condition or practice with *probable potential* for causing *nondisabling* injury or illness or *nondisruptive* property damage.

10.5 Unsafe Conditions Involving Mechanical or Physical Facilities*

The total working environment must be under constant scrutiny because of changing conditions, new employees, equipment additions and modifications, and so on. The following checklist is presented as a guide to identify potential problems:

1. Building

- Correct ceiling height
- Correct floor type; in acceptable condition
- Adequate illumination
- Adequate plumbing and heating pipes and equipment
- Windows with acceptable opening, closing, and holding devices; protection from breakage
- Acceptable size doors with correct swing and operational quality
- Adequate railing and nonslip treads on stairways and balconies
- Adequate ventilation
- Adequate storage facilities
- Adequate electrical distribution system in good condition
- Effective space allocation
- Adequate personal facilities (restrooms, drinking fountains, wash-up facilities, etc.)
- Efficient traffic flow
- Adequate functional emergency exits
- Effective alarms and communications systems
- Adequate fire prevention and extinguishing devices
- Acceptable interior color scheme
- Acceptable noise absorption factor
- Adequate maintenance and cleanliness

2. Machinery and equipment

- Acceptable placement, securing, and clearance
- Clearly marked safety zones
- Adequate nonskid flooring around machines
- Adequate guard devices on all pulleys
- Sharp, secure knives and cutting edges

* From J. B. ReVelle and J. Stephenson, *Safety Training Methods*, 2nd ed. Copyright © 1995. Reprinted by permission of John Wiley & Sons, New York.

Properly maintained and lubricated machines in good working condition
 Functional, guarded, magnetic-type switches on all major machines
 Properly wired and grounded machines
 Functional hand and portable power tools in good condition and grounded
 Quality machines adequate to handle the expected work load
 Conspicuously posted safety precautions and rules near each machine
 Guards for all pinch points within 7 ft of the floor

11 SAFETY TRAINING

11.1 Specialized Courses*

Automated External Defibrillator Training

Learning to use an automated external defibrillator (AED) is highly intuitive and surprisingly simple. Many people report that it is far easier than learning cardiopulmonary resuscitation (CPR) (*cardio* means “heart,” *pulmonary* means “lungs,” and *resuscitation* means “breathing for someone”). When a rescuer is performing CPR, he or she is attempting to bring the person back to life (resuscitate). Current AED courses last up to four hours to allow ample time for hands-on practice and to help increase user competence and confidence. AED training and related resources are offered through the American Heart Association, the American Red Cross, EMP America, the American Health and Safety Institute, the National Safety Council, and others. AED manufacturers also offer training courses. Since most states regulate health care training for public safety personnel, it is a good idea to check with your state authorities to make sure your training program is consistent with state guidelines. To do this, contact your state Emergency Medical Service (EMS) agency.

AED training curricula vary but generally emphasize:

- A working knowledge of CPR
- Safety for both victims and rescuers
- Proper placement of electrodes
- Delivering the first shock as quickly as possible, ideally within 60 s from time of arrival at the victim’s side
- Plenty of hands-on practice, with one instructor and one AED or AED trainer for every four to six students

Experience has shown that emergency responders may go for several years before encountering a victim in cardiac arrest. Lay rescuers may use an AED only once in a lifetime. Therefore, it is important to review AED skills on a regular basis. The ideal frequency for retraining is unknown, but most experts recommend reviewing AED skills every three to four months.

Driver Training

The number one accident killer of employees is the traffic accident. Each year more than 27,000 workers die in non-work-related motor vehicle accidents, and an additional 3900 employees are killed in work-related accidents. The employer pays a heavy toll for these accidents. Those that are work related are compensable, but the others are, nonetheless, costly. The loss of a highly

* From R. De Reamer, *Modern Safety and Health Technology*. Copyright © 1980. Reprinted by permission of John Wiley & sons, New York

skilled worker, a key scientist, or a company executive could have a serious impact on the success of the business.

There is, fortunately, something constructive that employers can do to help protect their employees and their executives from the tragedy and waste of traffic accidents. Driver training for workers and executives can be provided either in-house or through community training agencies. Companies that have conducted driver-training programs report that the benefits of such training were not limited to the area of improved traffic accident performance. These companies also experienced lower on-the-job injury frequency rates (the training produced an increase in safety awareness) and improved employee–community relations.

Companies have taken several approaches to driver training:

- A course has been made available to employees on a volunteer basis, either on- or off-hours.
- Driver training has been made mandatory for employees who operate a motor vehicle on company business.
- The company has promoted employee attendance at community–agency-operated programs.
- Full-scale driver-training programs have been conducted for all employees and members of their families. This is done off-hours and attendance is voluntary.

Environmental Risk Training

How cold or hot is it outside? Simply knowing the temperature does not tell you enough about the conditions of the working environment to enable an employee to know how to dress sensibly for weather extremes, whether hot or cold. Other factors, including wind speed, relative humidity, dew point, and cloud conditions, play important roles in determining how hot or cold it feels outside. Two indices, the heat index and the wind chill index, have been created to explain these conditions and employees whose work takes them outside should receive some practical training about the indices and how to best use them to avoid the otherwise painful and possibly damaging consequences of temperature extremes.

Heat Index. In an average year only the winter’s cold—not lightning, hurricanes, tornadoes, floods, or earthquakes—takes a greater weather-related death toll than the summer’s heat and humidity. In an effort to alert persons to the hazards of prolonged heat/humidity episodes, the National Weather Service devised the *heat index* (HI). The HI is an accurate measure of how hot it really feels when the effects of relative humidity are added to high temperature.

The human body contains several mechanisms to maintain its internal operating temperature at approximately 98.6°F. When threatened with above normal temperatures, the body will try to dissipate excess heat by varying the rate and depth of blood circulation, by losing water through the skin and sweat glands, and, as a last resort, by panting. When weather conditions force the air temperature above 90°F and the relative humidity is high, the body is doing everything it can to maintain normal temperature. Unfortunately, conditions can exceed the body’s ability to cope with the combined efforts of heat and humidity. At such times the body may succumb to any of a number of heat disorders, including sunstroke, heat cramps, heat exhaustion, and heatstroke.

To use the HI charts, find the appropriate temperature at the top of the chart. Read down until you are opposite the relative humidity or the dew point. The number that appears at the intersection of the temperature and humidity/dew point is the HI. Refer to Tables 13 and 14.

The HI is the “feels like” or apparent temperature. As relative humidity increases, the air seems warmer than it actually is because the body is less able to cool itself via evaporation of perspiration. As the HI rises, so do health risks. When the HI is between 90 and 105°F,

Table 13 Heat Index: Temperature and Relative Humidity

RH (%)	By Temperature (°F)															
	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
90	119	123	128	132	137	141	146	152	157	163	168	174	180	186	193	199
85	115	119	123	127	132	136	141	145	150	155	161	166	172	178	184	190
80	112	115	119	123	127	131	135	140	144	149	154	159	164	169	175	180
75	109	112	115	119	122	126	130	134	138	143	147	152	156	161	166	171
70	106	109	112	115	118	122	125	129	133	137	141	145	149	154	158	163
65	103	106	108	111	114	117	121	124	127	131	135	139	143	147	151	155
60	100	103	105	108	111	114	116	120	123	126	129	133	136	140	144	148
55	98	100	103	105	107	110	113	115	118	121	124	127	131	134	137	141
50	96	98	100	102	104	107	109	112	114	117	119	122	125	128	131	135
45	94	96	98	100	102	104	106	108	110	113	115	118	120	123	126	129
40	92	94	96	97	99	101	103	105	107	109	111	113	116	118	121	123
35	91	92	94	95	97	98	100	102	104	106	107	109	112	114	116	118
30	89	90	92	93	95	96	98	99	101	102	104	106	108	110	112	114

Note. Exposure to full sunshine can increase HI values by up to 15°F.

Table 14 Heat Index: Temperature and Dew Point

Dewpoint (°F)	By Temperature (°F)															
	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
65	94	95	96	97	98	100	101	102	103	104	106	107	108	109	110	112
66	94	95	97	98	99	100	101	103	104	105	106	108	109	110	111	112
67	95	96	97	98	100	101	102	103	105	106	107	108	110	111	112	113
68	95	97	98	99	100	102	103	104	105	107	108	109	110	112	113	114
69	96	97	99	100	101	103	104	105	106	108	109	110	111	113	114	115
70	97	98	99	101	102	103	105	106	107	109	110	111	112	114	115	116
71	98	99	100	102	103	104	106	107	108	109	111	112	113	115	116	117
72	98	100	101	103	104	105	107	108	109	111	112	113	114	116	117	118
73	99	101	102	103	105	106	108	109	110	112	113	114	116	117	118	119
74	100	102	103	104	106	107	109	110	111	113	114	115	117	118	119	121
75	101	103	104	106	107	108	110	111	113	114	115	117	118	119	121	122
76	102	104	105	107	108	110	111	112	114	115	117	118	119	121	122	123
77	103	105	106	108	109	111	112	114	115	117	118	119	121	122	124	125
78	105	106	108	109	111	112	114	115	117	118	119	121	122	124	125	126
79	106	107	109	111	112	114	115	117	118	120	121	122	124	125	127	128
80	107	109	110	112	114	115	117	118	120	121	123	124	126	127	128	130
81	109	110	112	114	115	117	118	120	121	123	124	126	127	129	130	132
82	110	112	114	115	117	118	120	122	123	125	126	128	129	131	132	133

Note. Exposure to full sunshine can increase HI values by up to 15°F.

heat exhaustion is possible. When it is above 105°F, it is probable. Heatstroke is possible when the HI is above 105°F and is quite likely when it is 130°F and above. Physical activity of any kind and prolonged exposure to the heat increase the risks. Environmental risk training for employees should also include this information:

- *Heat exhaustion* occurs when the body is dehydrated.
Symptoms. Headache, nausea, dizziness, cool and clammy skin, pale face, cramps, weakness, profuse perspiration.

First Aid. Move to a cooler spot, drink water with a small amount of salt added (one teaspoon per quart).

Result. It can lead to collapse and heatstroke.

- *Heatstroke* occurs when perspiration cannot occur and the body overheats.

Symptoms. Headache, nausea, face flushed, hot and dry skin, perspiration, body temperature over 101°F, chills, rapid pulse.

First Aid. Cool person immediately, move to shade or indoors, wrap in a cool, wet sheet, get medical assistance.

Result. It can lead to confusion, coma, and death.

Wind Chill Index. A description of the character of weather known as “coldness” was proposed around 1940 by scientists working in the Antarctic. The *wind chill index* was developed to describe the relative discomfort/danger resulting from the combination of wind and temperature. The wind chill index describes an equivalent temperature at which the heat loss from exposed flesh would be the same as if the wind were near calm. For example, a wind chill index of -5 indicates that the effects of wind and temperature on exposed flesh are the same as if the air temperature were 5 degrees below zero, even though the actual temperature is higher.

The high importance of the wind chill index to exposed persons is as an indicator of how to dress properly for winter weather. (Wind chill does not affect a car’s antifreeze protection, freezing of water pipes, etc.) In dressing for cold weather, an important factor to remember is that entrapped insulating air warmed by body heat is the best protection against the cold. Consequently, persons working outside should wear loose-fitting, lightweight, warm clothing in several layers. Outer garments should be tightly woven, water repellant, and hooded. Mittens snug at the wrist are better protection than fingered gloves. Physical activity of any kind and prolonged exposure to the cold increase the risks. As with the HI, environmental risk training for employees should include this information.

To use the wind chill index chart, find the approximate temperature at the top of the chart. Read down until you are opposite the appropriate wind speed. The number found at the intersection of the temperature and wind speed is the wind chill index. Refer to Table 15.

If the engineer in you would like to calculate the wind chill index for various combinations of temperature and wind speed other than those in Table 15, the following formula should be used:

$$WCI = 91.4 - [0.474677 - 0.020425V + 0.303107 \times \text{SQRT}(V)](91.4 - T)$$

Table 15 Wind Chill Index

Wind (mph)	By Temperature (°F)												
	35	30	25	20	15	10	5	0	-5	-10	-15	-20	-25
5	32	27	22	16	11	6	0	-5	-10	-15	-21	-26	-31
10	22	16	10	3	-3	-9	-15	-22	-27	-34	-40	-46	-52
15	16	9	2	-5	-11	-18	-25	-31	-38	-45	-51	-58	-65
20	12	4	-3	-10	-17	-24	-31	-39	-46	-53	-60	-67	-74
25	8	1	-7	-15	-22	-29	-36	-44	-51	-59	-66	-74	-81
30	6	-2	-10	-18	-25	-33	-41	-49	-56	-64	-71	-79	-86
35	4	-4	-12	-20	-27	-35	-43	-52	-58	-67	-74	-82	-92
40	3	-5	-13	-21	-29	-37	-45	-53	-60	-69	-76	-84	-92

Note. Wind speeds above 40 mph have little additional chilling effect. Values of wind chill below -10° F are considered bitterly cold. Values of wind chill below -20° F are extremely cold—human flesh will begin to freeze within 1 min!

where WCI = wind chill index

V = wind speed (miles per hour)

SQRT = square root

T = temperature ($^{\circ}$ F)

Fire Protection Training

All employees must know what to do when a fire alarm sounds. All employees must know something about the equipment provided for fire protection and what they can do toward preventing a fire. They must know:

- The plan established for evacuation of the building in case of emergency.
- How to use the first-aid fire appliances provided (extinguishers, hose, etc.).
- How to use other protective equipment. (All employees should know that water to extinguish fires comes out of the pipes of the sprinkler systems and that stock must not be piled so close to sprinkler lines that it prevents good distribution of water from sprinkler heads on a fire in the piled material. They should know that fire doors must be kept operative and not obstructed by stock piles, tools, or other objects.)
- How to give a fire alarm and how to operate plant fire alarm boxes and street boxes in the public alarm system.
- Where smoking in the plant is permitted and where, for fire safety reasons, it is prohibited.
- The housekeeping routine (disposal of wiping rags and waste, handling of packing materials, and other measures for orderliness and cleanliness throughout the plant).
- Hazards of any special process in which the employee is engaged.

All these “what-to-do” items can appropriately be covered in training sessions and evacuation drills.

First-Aid Training

First-aid courses pay big dividends in industry. This statement is based on clear evidence that people trained in first aid are more safety conscious and less likely to have an accident. The importance of first-aid training from the safety standpoint is that it teaches much more than applying a bandage or a splint. According to the Red Cross, “The primary purpose of first aid training is the prevention of accidents.” Each lesson teaches the student to analyze (1) how the accident happened, (2) how the accident could have been prevented, and (3) how to treat the injury. But the biggest dividend of first-aid training is the lives that have been saved because trainees were prepared to apply mouth-to-mouth resuscitation, to stop choking using the Heimlich maneuver (ejection of foreign object by forceful compression of diaphragm), or to stem the flow of blood.

Since the OSHAct, first-aid training has become a matter of federal law—the act stipulates that in the absence of an infirmary, clinic, or hospital in proximity to the workplace, a person or persons should be adequately trained to render first aid. The completion of the basic American National Red Cross first-aid course will be considered as having met this requirement. Just what constitutes *proximity* to a clinic or hospital? The OSH Review Commission, recognizing that first aid must be given within 3 minutes of serious accidents, concluded that an employer whose plant had no one trained in first aid present and was located 9 minutes from the nearest hospital violated the standard (1910.151, Medical and First Aid).

Hazard Recognition Training

The OSHAct of 1970 makes it quite clear that employers are solely responsible for supplying a work environment that is, insofar as is possible, free of hazards. Nonetheless, unavoidable

hazards can and do occur. It is incumbent on employers to develop or acquire as well as to deliver hazard recognition training to affected employees. The purpose of this training is to ensure that employees are fully knowledgeable and capable of recognizing workplace hazards. In this context workplace hazards can include a broad spectrum of situations that range from missing rubber feet on ladders to dangerous gases in confined spaces.

HAZWOPER Training

The Hazardous Waste Operations and Emergency Response Standard (HAZWOPER) applies to five distinct groups of employers and their employees. This includes any employees who are exposed or potentially exposed to hazardous substances, including hazardous waste, and who are engaged in one of the following operations as specified by 29 CFR 1910.120(a)(1)(iv) and 1926.65(a)(1)(i-v):

- Cleanup operations—required by a governmental body, whether federal, state, local, or other involving hazardous substances—that are conducted at uncontrolled hazardous waste sites
- Corrective actions involving cleanup operations at sites covered by the *Resource Conservation and Recovery Act of 1976 (RCRA)*, as amended
- Voluntary cleanup operations at sites recognized by federal, state, local, or other governmental body as uncontrolled hazardous waste sites
- Operations involving hazardous wastes that are conducted at treatment, storage, and disposal facilities regulated by 40 CFR, Parts 264 and 265 pursuant to RCRA, or by agencies under agreement with the EPA to implement RCRA regulations
- Emergency response operations for releases of or substantial threats of hazardous substances regardless of the location of the hazard

Computer-based training may meet some refresher training requirements, provided that it covers topics relevant to workers' assigned duties. It must be supplemented by the opportunity to ask questions of a qualified trainer and by an assessment of hands-on performance of work tasks. HAZWOPER refresher training may be given in segments so long as the required eight hours have been completed by the employee's anniversary date.

If the date for the refresher training has lapsed, the need to repeat initial training must be determined based on the employee's familiarity with safety and health procedures used on site. The employee should take the next available refresher training course. There should be a record in the employee's file indicating why the training has been delayed and when the training will be completed.

Lockout Box Training

The OSHA standard for *The Control of Hazardous Energy (Lockout/Tagout)*, 29 CFR 1910.147, addresses the practices and procedures necessary to disable machinery or equipment, thereby preventing the release of hazardous energy while employees perform servicing and maintenance activities. The standard outlines measures for controlling hazardous energies, i.e., electrical, mechanical, hydraulic, pneumatic, chemical, thermal, and other energy sources. In addition, 29 CFR 1910.333 sets forth requirements to protect employees working on electrical circuits and equipment. This section requires workers to use safe work practices, including lockout and tagging procedures. These provisions apply when employees are exposed to electrical hazards while working on, near, or with conductors or systems that use electric energy. The lockout/tagout standard establishes the employer's responsibility to protect employees from hazardous energy sources on machines and equipment during service and maintenance.

Employees need to be trained to ensure that they know, understand, and follow the applicable provisions of the hazardous energy control procedures. The training must cover at least three areas: aspects of the employer's energy control program, elements of the energy control procedure relevant to the employee's duties or assignment, and the various requirements of the OSHA standards related to lockout/tagout (LOTO).

Machine Guard Training

Moving machine parts have the potential to cause severe workplace injuries, such as crushed fingers or hands, amputations, burns, or blindness. Safeguards are essential for protecting workers from these needless and preventable injuries. Any machine part, function, or process that can cause injury must be safeguarded. When the operation of a machine or accidental contact with it can injure the operator or others in the vicinity, the hazards must be either eliminated or controlled.

Machine guard training should be provided to workers in advance of their assignments to use any machines that require the use of machine guards. The training should include exposure to the actual machines or, at the very least, multiple photographs of the machines. As much as is practical, the employee and trainer walkaround and/or the photographs should allow the machines to be viewed both with and without the required guards. Photographs that include machine operators should indicate the use of appropriate PPE.

The machine guard training should address the four types of guards:

1. Fixed guards
 - A permanent part of the machine
 - Not dependent on any other part to perform the function
 - Usually made of sheet metal, screen, bars, or other material that will withstand the anticipated impact
 - Generally considered the preferred type of guard
 - Simple and durable
2. Interlocked guards
 - Usually connected to a mechanism that will cut off power automatically
 - Could use electrical, mechanical, or hydraulic systems
 - Should rely on a manual reset system
3. Adjustable guards
 - Very flexible to accommodate various types of stock
 - Manually adjusted
4. Self-adjusting guards
 - Opening is determined by the movement of the stock through the guard
 - Does not always provide maximum protection
 - Common complaint is reduced visibility at the point of operation: "I can't see what I'm doing!"

Upon completion of machine guard training, operators should be able to immediately recognize the presence or absence of the correct type of machine guard for any machine they will be expected to operate. Furthermore, operators should be able to immediately detect if a machine guard is properly installed so as to provide maximum operator protection.

Personal Protection Equipment/Physical Protection Training

Every employee should receive whatever PPE that is needed to ensure short- and long-term safety and health. The following list of PPE is typical, but not exhaustive:

- Coveralls
- Earplugs/earmuffs
- Gloves
- Footwear (protective and high traction)
- Protective hoods
- Respirators
- Safety glasses
- Self-contained breathing apparatus (SCUBA)

The intent of supplying the correct PPE to the correct employee at the correct time in the correct condition is to ensure the necessary protection of their physical parts. The following list of physical parts is typical but not exhaustive:

- Back
- Ears
- Elbows
- Eyes
- Feet
- Hands
- Knees
- Mouth (unintended ingestion)
- Nose (unintended inhalation)

Other Specialized Courses

Some of the other specialized courses that can be given for safety training are:

- Accident investigation
- Accident report preparation
- Eye/face wash stations
- Powered equipment and vehicles
- Safety recordkeeping
- Specific disasters

11.2 Job Hazard Analysis Training*

Admittedly, the conventional mass approach to safety training takes little of the supervisor's time. Group training sessions, safety posters, films, and booklets are handled by the plant safety engineer, human resources, or other staff people. On the other hand, where safety training is carried out on a personalized basis, the first-line supervisor is the right person to provide the

* From R. De Reamer, *Modern Safety and Health Technology*. Copyright © 1980. Reprinted by permission of John Wiley & Sons, New York.

Table 16 Job Hazard Analysis Form

Job description: Three-spindle drill press—Impeller 34C6

Job location: *Bldg. 19-2, Pump Section*

Key Job Steps	Tool Used	Potential Health and Injury Hazard	Safe Practices, Apparel, and Equipment
Get material from operation	Tote box	Dropping tote box on foot	Wear safety shoes. Have firm grip on box.
Inspect and set up drill press	Drill press	Back strain from lifting; picking up overloaded boxes Check for defective machines Chuck wrench not removed	Stress proper lifting methods. Tell employee to get help or lighten load. Do not operate if defective.
Drilling		Chuck wrench not removed Making adjustments when machine is running Hair, clothing, or jewelry catching on spindle Spinning work or fixture Injury to hands—cuts, etc. Drill sticks in work Flying chips Pinch points at belts Broken drills James Black Signature _____ <u>7/22/05</u> Date	Always remove chuck wrench immediately after use. Always stop spindle before making adjustments. Wear head covering, snug-fitting; clothing. No loose sleeves. Avoid wearing rings, bracelets, or wristwatches. Use proper blocks or clamps to hold work and fixture securely. Never wear gloves. Use hook, brush, or other tool to remove chips. Use compressed air only when instructed. Stop spindle, free drill by hand. Wear proper eye protection. Always stop press before adjusting belts. Do not attempt to force drill, apply pressure.
		<u>1 of 3</u> Page	

training. This will take more of his or her time and require more attention to detail, but this additional effort pays off because of the increased effectiveness of the training method.

In launching a personalized safety training program, the first step is the preparation of a job hazard analysis for each job in the plant. To make the job hazard analysis in an organized manner, use of a form similar to the one shown in Table 16 is suggested. The key elements of the form are (1) job description; (2) job location; (3) key job steps; (4) tool used; (5) potential health and injury hazards; and (6) safe practices, apparel, and equipment.

A review of the form will indicate the steps in making a job hazard analysis. To start an analysis, the key steps of the job are listed in order in the first column of the form. Where pertinent, the tool used to perform the job step is listed in the second column. In the third column opposite each job step, the hazards of the particular step are indicated. Finally, in the fourth column of the form are listed the safe practices that the employee must be shown and have discussed. Here the supervisor lists the safe work habits that must be stressed and the safety equipment and clothing required for the job.

In making the analysis, an organized approach is required so the less obvious accident hazards will not be missed. This means going out on the floor and actually watching the job being performed and jotting down key steps and hazards as they are observed. Supervisors who make such a job hazard analysis are often surprised to find hazards in the job cycle that they had missed seeing in the past. Their original negative reaction to the thought of additional paperwork soon disappears. In the long run, supervisors realize that proper hazard analysis will help them do a better training job.

As previously stated, a job hazard analysis is made for each job. In most cases, each supervisor will have to make from 5 to 10 different analyses. Of course, in maintenance and

construction work, the variety of jobs covers a much wider range. Fortunately, these jobs can be grouped by the type of work performed and a job hazard analysis can be made for each category of work, rather than for each job. For example, repair, installation, and relocation of equipment; cleaning motors; and unloading cars might be a few of the various categories of maintenance work to be analyzed.

Some of the topics that may need to be addressed as part of job hazard analysis training include:

- Air-sampling/gas-monitoring equipment (for such compounds as ammonia, ethyl/methyl Mercapton, formaldehyde, hydrogen sulfide, and sulfur dioxide)
- HAZMAT kits
- Odor recognition (for the same compounds)
- Oil immobilants
- Personal skill responses
- Sound-level meters

11.3 Management's Overview of Training*

An effective accident prevention program requires proper job performance from everyone in the workplace. All employees must know about the materials and equipment they are working with, what hazards are known in the operation, and how these hazards have been controlled or eliminated. Each individual employee needs to know and understand the following points (especially if they have been included in the company policy and in a "code of safe practices"):

- No employee is expected to undertake a job until he or she has received instruction on how to properly perform the job and has been authorized to perform that job.
- No employee should undertake a job that appears to be unsafe.
- Mechanical safeguards are in place and must be kept in place.
- Each employee is expected to report all unsafe conditions encountered during work.
- Even slight injury or illness suffered by an employee must be reported at once.

In addition to the points above, any safety rules that are a condition of employment, such as the use of safety shoes or eye protection, should be explained clearly and enforced at once.

The first-line supervisors must know how to train employees in the proper way of doing their jobs. Encourage and consider providing for supervisory training for these supervisors. (Many colleges offer appropriate introductory management training courses.)

Some specific training requirements in the OSHA standards must be met, such as those that pertain to first aid and powered industrial trucks (including forklifts). In general, they deal with situations where the use of untrained or improperly trained operators on skill machinery could cause hazardous situations to develop, not only for the operator, but possibly for nearby workers, too.

Particular attention must be given to new employees. Immediately on arriving at work, new employees begin to learn things and to form attitudes about the company, the job, their boss, and their fellow employees. Learning and attitude formation occur regardless of whether the employer makes a training effort. If the new employees are trained during those first few hours and days to do things the right way, considerable losses may be avoided later.

* From J. B. ReVelle and J. Stephenson. *Safety Training Methods*, 2nd ed. Copyright © 1995. Reprinted by permission of John Wiley & Sons, New York.

At the same time, attention must be paid to regular employees, including the old-timers. Old habits can be wrong habits. An employee who continues to repeat an unsafe procedure is not working safely, even if an “accident” has not resulted from this behavior.

Although every employee’s attitude should be one of determination that “accidents” can be prevented, one thing more may be needed. It should be stressed that the responsibility assigned to the person in charge of the job—as well as to all other supervisors—is to be sure that there is a concerted effort under way at all times to follow every safe work procedure and health practice applicable to that job. It should be clearly explained to these supervisors that they should never silently condone unsafe or unhealthful activity in or around any workplace.

11.4 Sources and Types of Training Materials

Sources of training materials include OSHA, the Department of Transportation (DOT), the EPA, and HAZWOPER. Types of training materials include printed documents (books, manuals, and pamphlets), videos (tapes and compact discs), and the Internet. The bibliography to this chapter presents an extensive list of sources in a variety of types of training.

BIBLIOGRAPHY

- “Accident Prevention: Your Key to Controlling Surging Workers’ Compensation Costs,” *Occupational Hazards*, 35 (November 1979).
- “Accident Related Losses Make Cost Soar,” *Industrial Engineering*, 26 (May 1979).
- “Analyzing a Plant Energy-Management Program, Part I: Measuring Performance,” *Plant Engineering*, 59 (October 30, 1980).
- “Analyzing a Plant Energy-Management Program, Part II: Forecasting Consumption,” *Plant Engineering*, 149 (November 13, 1980).
- “Anatomy of a Vigorous In-Plant Program,” *Occupational Hazards*, 32 (July 1979).
- “A Shift Toward Protective Gear,” *Business Week*, 56H (April 13, 1981).
- “A Win for OSHA,” *Business Week*, 62 (June 29, 1981).
- “Buyers Should Get Set for Tougher Safety Rules,” *Purchasing*, 34 (May 25, 1976).
- “Complying with Toxic and Hazardous Substances Regulations, Part I,” *Plant Engineering*, 283 (March 6, 1980).
- “Complying with Toxic and Hazardous Substances Regulations, Part II,” *Plant Engineering*, 157 (April 17, 1980).
- “Computers Help Pinpoint Worker Exposure,” *Chemecology*, 11 (May 1981).
- “Conserving Energy by Recirculating Air from Dust Collection Systems,” *Plant Engineering*, 151 (April 17, 1980).
- “Control Charts Help Set Firm’s Energy Management Goals,” *Industrial Engineering*, 56 (December 1980).
- “Controlling Noise and Reverberation with Acoustical Baffles,” *Plant Engineering*, 131 (April 17, 1980).
- “Controlling Plant Noise Levels,” *Plant Engineering*, 127 (June 24, 1976).
- “Cost–Benefit Decision Jars OSHA Reform,” *Industry Week*, 18 (June 29, 1981).
- “Cost Factors for Justifying Projects,” *Plant Engineering*, 145 (October 16, 1980).
- “Costs, Benefits, Effectiveness, and Safety: Setting the Record Straight,” *Professional Safety*, 28 (August 1975).
- “Costs Can Be Cut Through Safety,” *Professional Safety*, 34 (October 1976).
- “Cutting Your Energy Costs,” *Industry Week*, 43 (February 23, 1981).
- De Reamer, R., *Modern Safety and Health Technology*, Wiley, New York, 1980.
- “Elements of Effective Hearing Protection,” *Plant Engineering*, 203 (January 22, 1981).
- “Energy Constraints and Computer Power Will Greatly Impact Automated Factories in the Year 2000,” *Industrial Engineering*, 34 (November 1980).
- “Energy Managers Gain Power,” *Industry Week*, 62 (March 17, 1980).

- “Energy Perspective for the Future,” *Industry Week*, 67 (May 26, 1980).
- Engineering Control Technology Assessment for the Plastics and Resins Industry*, NIOSH Research Report Publication No. 78-159.
- “Engineering Project Planner, A Way to Engineer Out Unsafe Conditions,” *Professional Safety*, 16 (November 1976).
- “EPA Gears Up to Control Toxic Substances,” *Occupational Hazards*, 68 (May 1977).
- Ferry, T. S., *Safety Management Planning*, The Merritt Company, Santa Monica, CA, 1982.
- “Fume Incinerators for Air Pollution Control,” *Plant Engineering*, 108 (November 13, 1980).
- “Groping for a Scientific Assessment of Risk,” *Business Week*, 120J (October 20, 1980).
- “Hand and Body Protection: Vital to Safety Success,” *Occupational Hazards*, 31 (February 1979).
- “Hazardous Wastes: Coping with a National Health Menace,” *Occupational Hazards*, 56 (October 1979).
- “Hearing Conservation—Implementing an Effective Program,” *Professional Safety*, 21 (October 1978).
- “How Do You Know Your Hazard Control Program Is Effective?” *Professional Safety*, 18 (June 1981).
- “How to Control Noise,” *Plant Engineering*, 90 (October 5, 1972).
- “Human Factors Engineering—A Neglected Art,” *Professional Safety*, 40 (March 1978).
- “IE Practices Need Reevaluation Due to Energy Trends,” *Industrial Engineering*, 52 (December 1980).
- “Industrial Robots: A Primer on the Present Technology,” *Industrial Engineering*, 54 (November 1980).
- “Job-Safety Equipment Comes Under Fire, Are Hard Hats a Solution or a Problem?” *The Wall Street Journal*, 40 (November 18, 1977).
- Johnson, W. G., *MORT Safety Assurance Systems*, Marcel Dekker, New York, 1979.
- McFarland, R. A., “Application of Human Factors Engineering to Safety Engineering Problems,” *National Safety Congress Transactions*, 12, National Safety Council, Chicago, 1967.
- “New OSHA Focus Led to Noise-Rule Delay,” *Industry Week*, 13 (June 15, 1981).
- “OSHA Communiqué,” *Occupational Hazards*, 27 (June 1981).
- “OSHA Moves Health to Front Burner,” *Purchasing*, 46 (September 26, 1979).
- “OSHA to Analyze Costs, Benefits of Lead Standard,” *Occupational Health & Safety*, 13 (June 1981).
- Patty's Industrial Hygiene and Toxicology*, Vol. 1, 3rd rev. ed., Wiley-Interscience, New York, 1978.
- “Practical Applications of Biomechanics in the Workplace,” *Professional Safety*, 34 (July 1975).
- “Private Sector Steps Up War on Welding Hazards,” *Occupational Hazards*, 50 (June 1981).
- “Putting Together a Cost Improvement Program,” *Industrial Engineering*, 16 (December 1979).
- “Reduce Waste Energy with Load Controls,” *Industrial Engineering*, 23 (July 1979).
- “Reducing Noise Protects Employee Hearing,” *Chemecology*, 9 (May 1981).
- “Regulatory Relief Has Its Pitfalls, Too,” *Industry Week*, 31 (June 29, 1981).
- ReVelle, J. B., and J. Stephenson, *Safety Training Methods*, 2nd ed., Wiley, New York, 1995.
- “ROI Analysis for Cost-Reduction Projects,” *Plant Engineering*, 109 (May 15, 1980).
- “Safety & Profitability—Hand in Hand,” *Professional Safety*, 36 (March 1978).
- “Safety Managers Must Relate to Top Management on Their Terms,” *Professional Safety*, 22 (November 1976).
- “Superfund Law Spurs Cleanup of Abandoned Sites,” *Occupational Hazards*, 67 (April 1981).
- “Taming Coal Dust Emissions,” *Plant Engineering*, 123 (May 15, 1980).
- “The Cost–Benefit Argument—Is the Emphasis Shifting?” *Occupational Hazards*, 55 (February 1980).
- “The Cost/Benefit Factor in Safety Decisions,” *Professional Safety*, 17 (November 1978).
- “The Design of Manual Handling Tasks,” *Professional Safety*, 18 (March 1980).
- “The Economics of Safety—A Review of the Literature and Perspective,” *Professional Safety*, 31 (December 1977).
- “The Hidden Cost of Accidents,” *Professional Safety*, 36 (December 1975).
- “The Human Element in Safe Man-Machine Systems,” *Professional Safety*, 27 (March 1981).
- “The Problem of Manual Materials Handling,” *Professional Safety*, 28 (April 1976).
- “Time for Decisions on Hazardous Waste,” *Industry Week*, 51 (June 15, 1981).
- “Tips for Gaining Acceptance of a Personal Protective Equipment Program,” *Professional Safety*, 20 (March 1976).

- “Toxic Substances Control Act,” *Professional Safety*, 25 (December 1976).
- “TSCA: Landmark Legislation for Control of Chemical Hazards,” *Occupational Hazards*, 79 (May 1977).
- “Were Engineering Controls ‘Economically Feasible’?” *Occupational Hazards*, 27 (January 1981).
- “Were Noise Controls ‘Technologically Feasible’?” *Occupational Hazards*, 37 (January 1981).
- “What Are Accidents Really Costing You?” *Occupational Hazards*, 41 (March 1979).
- “What’s Being Done About Hazardous Wastes?” *Occupational Hazards*, 63 (April 1981).
- “Where OSHA Stands on Cost-Benefit Analysis,” *Occupational Hazards*, 49 (November 1980).
- “Worker Attitudes and Perceptions of Safety,” *Professional Safety*, 28 (December 1981); 20 (January 1982).

CHAPTER 25

WHAT THE LAW REQUIRES OF THE ENGINEER

Alvin S. Weinstein and Martin S. Chizek
Weinstein Associates International
Delray Beach, Florida

1 ART OF THE ENGINEER	749	5.2 Hazard Index	763
1.1 Modeling for Real World	749	5.3 Design Hierarchy	763
1.2 Safety Factor	750	6 DEFENSE TO PRODUCT LIABILITY	764
2 PROFESSIONAL LIABILITY	751	6.1 State of the Art	764
2.1 Liability of Employee	751	6.2 Role of Safety Standards	765
2.2 Liability of Business	755	6.3 Contributory/Comparative Negligence	766
3 LAWS OF PRODUCT LIABILITY	757	6.4 Assumption of Risk	767
3.1 Definition	757	7 RECALLS, RETROFITS, AND CONTINUING DUTY TO WARN	768
3.2 Negligence	758	7.1 After-Market Hazard Recognition	768
3.3 Strict Liability	758	7.2 Types of Corrective Action	769
3.4 Express Warranty and Misrepresentation	759	8 DOCUMENTATION OF DESIGN PROCESS	770
4 NATURE OF PRODUCT DEFECTS	760	9 FINAL WORD	771
4.1 Production or Manufacturing Flaws	760		
4.2 Design Flaws	760		
4.3 Instructions and Warnings	761		
5 UNCOVERING PRODUCT DEFECTS	762		
5.1 Hazard Analysis	762		

1 ART OF THE ENGINEER

1.1 Modeling for Real World

Engineers believe that they practice their craft in a world of certainty. Nothing could be further from the truth! Because this chapter deals with the interface between law and technology, and because product liability is likely to be the legal area of concern to the engineer, our principal focus is on the engineering (design) of products or components of products.

Think for a moment about the usual way an engineer proceeds from a product concept to the resulting device. The engineer generally begins the design process with some type of specifications for the eventual device to meet, such as performance parameters, functional capabilities, size, weight, cost, and so on. Implicit, if not explicit, in the specifications are assumptions about the device's ultimate interaction with the real world. If the specification concerns, for example, loading or power needs that the device is either to produce or to withstand, someone has created

boundaries within which the product is to function. Clearly, there are bound to be some uncertainties, despite the specifying of precise values for the designers to meet. Even assuming that a given loading for a certain component is known with precision and repeatability, the design of the component more than likely will involve various assumptions: *how* the loading acts (e.g., point load or distributed); *when* it acts (e.g., static or dynamic); *where* it acts (e.g., two or three dimensions); and *what* it acts on (e.g., how sophisticated an analysis technique to use).

The point is that even with sophisticated and powerful computational tools and techniques, the real world is always modeled into one that could be analyzed and, as a result, is truly artificial. That is, a measure of uncertainty will always exist in any result, whatever the computational power. The question that is often unanswered or ignored in the design process is: How *much* uncertainty is there about the subtleties and exigencies of the true behavior of the environment (including people) on the product and the uncertainties in our, yes, artificial modeling technique?

1.2 Safety Factor

To mask the uncertainties and, frankly, to admit that, despite our avowal, the world from which we derive our design is not real but artificial, we incorporate a “safety factor.” Truly, it should be viewed as a factor of ignorance. We use it in an attempt to reestablish the real world from the one we have modified and simplified by our assumptions and to make it tractable, that is, so we can meet the product specifications. The function of the safety factor is to bridge the gap between the computational world and the one in which the product must actually function.

There are, in general, three considerations that are to be incorporated into the safety factor:

1. Uncertainties in material properties
2. Uncertainties in quality assurance
3. Uncertainties in the interaction of persons and the product—from the legal perspective, the most important of all

Example 1 Truck-Mounted Crane.

Consider a truck-mounted crane whose design specification is that it is to be capable of lifting 30 tons. The intent is, of course, that only under certain specific conditions, i.e., the boom angle, boom extension, rotational location of boom, etc., will the crane be able to lift 30 tons.

Inherent in the design, however, must be a safety factor cushion, not only to account for, e.g., the uncertainties in the yield stress of the steel or the possibility of some welds not being full penetration during fabrication but also for the uncertainties of the crane operator not knowing the precise weight of the load. In the real world, it is foreseeable that there will be times when no one on the job site knows or has ready access to sufficient data to know with reasonable certainty the weight of the load to be lifted.

The dilemma for the engineer–designer is how much latitude to allow in the load-lifting capability of the crane to accommodate uncertainty in the load weight. That is, the third component of the safety factor must reflect a realistic assessment of real-world uncertainties. The difficulty, of course, is that there are serious competing trade-offs to be considered in deciding on this element of the safety factor. For each percent above the 30-ton load specification that the engineer builds into the safety factor, the crane is likely to be heavier, larger, perhaps less maneuverable, etc. That is, the utility of the crane is likely to be increasingly compromised in one or more ways as the safety factor is increased.

Yet the engineer’s creed requires that the product must function in its true environment of use and do so with reasonable safety and reliability. The art of the engineer, then, is to balance competing trade-offs in design decision making to minimize the existence of hazards, while

acknowledging and accounting for human frailties, reasonably foreseeable product uses and misuses, and the true environment of product use.

And that is what the law requires of the engineer as well. We will explore some of these considerations later in this chapter. But first, we look at the issues of professional liability.

2 PROFESSIONAL LIABILITY

Whether engaged in research, development, manufacturing, engineering services, or technical consulting, today's engineer must be cognizant that the law imposes substantial accountability on both individual engineers and technology-related companies. Engineers can never expect to insulate themselves entirely from legal liability. However, they can limit their liability by maintaining a fundamental understanding of the legal concepts they are likely to encounter in the course of their career, such as professional negligence, agency, employment agreements, intellectual property rights, contractual obligations, and liability insurance.

2.1 Liability of Employee

Negligence and Standard of Care

A lawsuit begins when a person (corporations, as well, are considered as "persons" for legal purposes) whose body or property is injured or damaged alleges that the injury was caused by the acts of another and files a complaint. The person asserting the complaint is the *plaintiff*; the person against whom the complaint is brought is the *defendant*.

In the complaint, the plaintiff must state a *cause of action* (a legal theory or principle) that would, if proven to the satisfaction of the jury, permit the plaintiff to recover damages. If the cause of action asserted is *negligence*, then the plaintiff must prove, first, that the defendant owed the plaintiff a *duty* (i.e., had a responsibility toward the plaintiff). Then the plaintiff must show that the defendant *breached* that duty and, consequently, that the breach of duty by the defendant was the *cause* of the plaintiff's injury.

The doctrine of negligence rests on the duty of every person to exercise due care in his or her conduct toward others. A breach of this duty of care that results in injury to persons or property may result in a *tort* claim, which is a civil wrong (as opposed to a criminal wrong) for which the legal system compensates the successful plaintiff by awarding money damages. To make out a cause of action in negligence, it is not necessary for the plaintiff to establish that the defendant either intended harm or acted recklessly in bringing about the harm. Rather, the plaintiff must show that the defendant's actions fell below the *standard of care* established by law.

In general, the standard of care that must be exercised is that conduct which the *average reasonable person* of ordinary prudence would follow under the same or similar circumstances. The standard of care is an external and objective one and has nothing to do with individual subjective judgment, though higher duties may be imposed by specific statutory provisions or by reason of special knowledge.

Example 2 Negligent or Not?

Suppose a person is running down the street knocking people aside and causing injuries. Is this person breaching the duty to care to society and acting negligently? To determine this, we need to undertake a risk/utility analysis, i.e., does the utility of the action outweigh the harm caused?

If this person is running to catch the last bus to work, then the risk to others as he pushed them aside probably outweighs the utility to this person, which is getting to work on time. However, if the person has seen a knife-wielding assailant attacking someone and is trying to reach the policeman on the corner, then the utility (saving human life) is great. In such a case,

perhaps society should allow the possible harm caused to others as he pushed them aside to reach the police officer and thus not find the person negligent, even though others persons were injured in the attempt to reach the police officer.

No duty is imposed on a person to take precautions against events that cannot reasonably be foreseen. However, the professional must utilize such superior judgment, skill, and knowledge as he or she actually possesses. Thus, the professional mechanical engineer might be held liable for miscalculating the load-lifting capability in the previous crane example, while a general engineering technician might not.

The duty to exercise reasonable care and avoid negligence does not mean that engineers guarantee the results of their professional efforts. Indeed, if an engineer can show that everything a reasonably prudent engineer might do was, in fact, done correctly, then liability cannot attach, even though there was a failure of some kind that caused injury.

Example 3 Collapse of a Reasonably Designed Overpass.

A highway overpass, when designed, utilized all of the acceptable analysis techniques and incorporated all of the features that were considered to be appropriate for earthquake resistance at that time. Years later, the overpass collapses when subjected to an earthquake of moderate intensity. At the time of the collapse, there are newer techniques and features that, in all likelihood, would have prevented the collapse had they been incorporated into the design.

It is unlikely that liability would attach to the engineers who created the original design and specifications as long as they utilized techniques that were reasonable at that time.

Additionally, liability depends on a showing that the negligence of the engineer was the direct and proximate cause of the damages. If it can be shown that there were other superseding causes responsible for the damages, the engineer may escape liability even though his or her actions deviated from professional standards.

Example 4 Collapse of a Negligently Designed Overpass.

Suppose, instead, that after the collapse of the overpass in the preceding example, a review of the original analysis conducted by the engineers reveals several deficiencies in critical specifications that reasonably prudent engineers would not have overlooked. However, the intensity of the earthquake was of such a magnitude that, with reasonable certainty, the overpass would have collapsed even if it had been designed using the appropriate specifications. The engineers, in this scenario, are likely to escape liability.

However, the law does allow “joint and severable” liability against multiple parties who act either in concert or independently to cause injury to a plaintiff. Other defenses to an allegation of negligence include the “state of the art” argument, contributory/comparative negligence, and assumption of the risk. These are discussed in Section 6.

An employer is generally liable for the negligence, carelessness, errors, and omissions of its employees. However, as we see in the next section, liability may attach to the engineer employee under the law of agency.

Agency and Authority

Agency is generally defined as the relationship that arises when one person (the principal) manifests an intention that another person (the agent) shall act on his behalf. A principal may appoint an agent to do any act except an act that, by its nature or by contract, requires personal performance by the principal. An engineer employee may act as an agent of his employer, just as an engineering consultant may act as an agent of her client.

The agent, of course, has whatever duties are expressly stated in the contract with the principal. Additionally, in the absence of anything contrary in the agreement, the agent has three major duties implied by law:

1. The fiduciary duty of an agent to his principal is one of undivided loyalty, e.g., no self-dealing or obtaining secret profits.
2. An agent must obey all reasonable directions of the principal.
3. An agent owes a duty to the principal to carry out his duties with reasonable care, in light of local community standards and taking into account any special skills of the agent.

Just as the agent has duties, the principal owes the agent a duty to compensate the agent reasonably for his services, indemnify the agent for all expenses or losses reasonably incurred in discharging any authorized duties, and, of course, comply with the terms of any contract with the agent.

With regard to tort liability in the context of the employer–employee relationship, an employer is liable only for those torts committed by an employee; he or she is not generally liable for torts committed by an agent functioning as an independent contractor. An example of an employee is one who works full-time, is compensated on a time basis, and is subject to the supervision of the principal in the details of his or her work. An example of an independent contractor is one who has a calling of his or her own, is hired to perform a particular job, is paid a given amount for that job, and follows his or her own discretion in carrying out the job. Engineering consultants are usually considered to be independent contractors.

Even when the employer–employee relationship is established, however, the employer is not liable for the torts of an employee unless the employee was acting within the scope of, or incidental to, the employer’s business. Additionally, the employer is usually not liable for the intentional torts of an employee on the simple ground that an intentional tort (i.e., fraud) is clearly outside the scope of employment. However, where the employee intentionally chooses a wrongful means to promote the employer’s business, such as fraud or misrepresentation, the employer may be held liable.

With regard to contractual liability under the law of agency, a principal will be bound on a contract that an agent enters into on his behalf if that agent has *actual authority*, i.e., authority expressly or implicitly contained within the agency agreement. The agent cannot be held liable to the principal for breach since he acted within the scope of his authority. To ensure knowledge of actual authority, the engineer should always obtain clear, written evidence of his job description, duties, responsibilities, “sign-off” authority, and so on.

Even where employment or agency actually exists, unless it is unequivocally clear that the individual engineer is acting on behalf of an employer or other disclosed principal, an injured third party has the right to proceed against either the engineer or the employer/principal or both under the rule that an agent for an undisclosed or partially disclosed principal is liable on the transaction together with her principal. Thus, engineers acting as employees or agents should always include their title, authority, and the name of the employer/principal when signing any contract or business document.

Even if the agent lacks actual authority, the principal can still be held liable on contracts entered into on his behalf if the agent had *apparent authority*, that is, where a third party reasonably believed, based on the circumstances, that the agent possessed actual authority to perform the acts in question. In this case, however, the agent may be held liable for losses incurred by the principal for unauthorized acts conducted outside the scope of the agent’s authority.

Employment Agreements

Rather than relying entirely on the law of agency to control the employer–employee relationship, most employers require engineers to sign a variety of employment agreements as a condition of employment. These agreements are generally valid and legally enforceable to the extent that they are reasonable in duration and scope.

A clause typically found in an engineer’s employment contract is the agreement of the employee to transfer the entire right, title, and interest in and to all ideas, innovations, and

creations to the company. These generally include designs, developments, inventions, improvements, trade secrets, discoveries, writings, and other works, including software, databases, and other computer-related products and processes. As long as the work is within the scope of the company's business, research, or investigation or the work resulted from or is suggested by any of the work performed for the company, its ownership is required to be assigned to the company.

Another common employment agreement is a noncompetition provision whereby the engineer agrees not to compete during his or her employment by the company and for some period after leaving the company's employ. These are also enforceable as long as the scope of the exclusion is reasonable in time and distance, when taking the nature of the product or service into account and the relative status of the employee. For example, courts would likely find invalid a two-year, nationwide noncompetition agreement against a junior computer-aided design/manufacturing (CAD/CAM) engineer in a small company; however, this agreement might be found fully enforceable against the chief design engineer of a large aircraft manufacturer. In any case, engineers should inform new/prospective employers of any prior employment agreement that is still in effect.

As will be seen in the next section, however, even if an employment agreement was not executed, ex-employees are not free to disclose or utilize proprietary information gained from their previous employers.

Intellectual Property

A *patent* is a legally recognized and enforceable property right for the exclusive use, manufacture, or sale of an invention by its inventor (or heirs or assignees) for a limited period of time that is granted by the government. In the United States, exclusive control of the invention is granted for a period of 20 years from the date of filing the patent and in consideration for which the right to free and unrestricted use passes to the general public. Patents may be granted to one or more individuals for *new* and *useful* processes, machines, manufacturing techniques, and materials, including improvements that are not obvious to one skilled in the particular art. The inventor, in turn, may license, sell, or assign patent rights to a third party. Remedies against patent infringers include monetary damages and injunctions against further infringement.

Engineers working with potentially patentable technology must follow certain formalities in the documentation and publication of information relating to the technology in order to preserve patent protection. Conversely, engineers or companies considering marketing a newly developed product or technology should have a patentability search conducted to ensure that they are not infringing existing patents.

Many companies rely on *trade secrets* to protect their technical processes and products. A trade secret is any information, design, device, process, composition, technique, or formula that is not known generally and that affords its owner a competitive business advantage. Advantages of trade secret protection include avoiding the cost and effort involved in patenting and the possibility of perpetual protection. The main disadvantage of a trade secret is that protection vanishes when the public is able to discover the "secret," whether by inspection, analysis, or reverse engineering. Trade secret protection thus lends itself more readily to intangible "know-how" than to end products.

Trade secrets have legal status and are protected by state common law. In some states, the illegal disclosure of trade secrets is classified as fraud, and employees can be fined or even jailed for such activity. Customer lists, supplier's identities, equipment, and plant layouts cannot be patented, yet they can be important in the conduct of a business and therefore are candidates for protection as trade secrets.

2.2 Liability of Business

Negligence for Services

Negligence (as defined in Section 2.1) and standards of care apply not only to individual engineers but also to consulting and engineering firms. At least one State Supreme Court has defined the standard of care for engineering services as follows:

In performing professional services for a client, an engineer has the duty to have that degree of learning and skill ordinarily possessed by reputable engineers, practicing in the same or a similar locality and under similar circumstances. It is his further duty to use the care and skill ordinarily used in like cases by reputable members of his profession practicing in the same or a similar locality, under similar circumstances, and to use reasonable diligence and his best judgment in the exercise of his professional skills and in the application of his learning, in an effort to accomplish the purpose for which he was employed.*

Occasionally, an engineer's duty to the general public may supersede the duty to her client. For example, an engineer retained to investigate the integrity of a building and determined the building was at imminent risk of collapse would have a duty to warn the occupants even if the owner requested that the engineer treat the results of the investigation as confidential.† The engineer also has a duty to adhere to applicable state and federal safety requirements. For example, the U.S. Department of Labor Occupational Safety and Health Administration has established safety and health standards for subjects ranging from the required thickness of a worker's hardhat to the maximum decibel noise level in a plant. In many jurisdictions, the violation of a safety code, standard, or statute that results in injury is "negligence per se," that is, a conclusive presumption of duty and breach of duty. Engineers should be aware, however, that the reverse of this rule does not hold true: Compliance with required safety standards does not necessarily establish reasonable care. See Section 6.2 for the role of safety standards.

Contractual Obligations

A viable contract, whether it be a simple purchase order to a vendor or a complex joint venture, requires the development of a working agreement that is mutually acceptable to both parties. An agreement (contract) binds each of the parties to do something or perhaps even refrain from doing something. As part of such an agreement, each of the parties acquires a legally enforceable right to the fulfillment of the promises made by the other. Breach of the contract may result in a court awarding damages for losses sustained by the nonbreaching party or requiring "specific performance" of the contract by the breaching party.

An oral contract can constitute just as binding a commitment as a written contract, although, by statute, some types of contracts are required to be in writing. As a practical matter, agreements of any importance should always be, and generally are, reduced to writing. However, a contract may also be created by implication based on the conduct of one party toward another.

In general, a contract must embody certain key elements, including (a) mutual assent as consisting of an offer and its acceptance between competent parties based on (b) valid consideration for (c) a lawful purpose or object in (d) clear-cut terms. In the absence of any one of these elements, a contract will generally not exist and hence will not be enforceable in a court of law.

Mutual assent is often referred to as a "meeting of the minds." The process by which parties reach this meeting of the minds generally is some form of negotiation, during which, at some

* *Clark v. City of Seward*, 659 P.2d 1227 (Alaska, 1983).

† California Attorney General's Opinion, Opinion No. 85-208 (1985).

point one party makes a proposal (offer) and the other agrees to it (acceptance). A counteroffer has the same effect as a rejection of the original offer.

To have a legally enforceable contract, there must generally be a bargained-for exchange of “consideration” between the parties, that is, a benefit received by the promisor or a detriment incurred by the promisee. The element of bargain assures that, at least when the contract is formed, both parties see an advantage in contracting for the anticipated performance.

If the subject matter of a contract (either the consideration or the object of a contract) is illegal, then the contract is void and unenforceable. Generally, illegal agreements are classified as such either because they are expressly prohibited by law (e.g., contracts in restraint of trade) or because they violate public policy (e.g., contracts to defraud others).

Problems with contracts can occur when the contract terms are incomplete, ambiguous, or susceptible to more than one interpretation or where there are contemporaneous conflicting agreements. In these cases, courts may allow other oral or written evidence to vary the terms of the contract.

A party that breaches a contract may be liable to the nonbreaching party for “expectation” damages, that is, sufficient damages to buy substitute performance. The breaching party may also be liable for any reasonably foreseeable consequential damages resulting from the breach.

Contract law generally permits claims to be made under a contract only by those who are “in privity,” that is, those parties among whom a contractual relationship actually exists. However, when a third party is an intended beneficiary of the contract or when contractual rights or duties have been transferred to a third party, then that third party may also have certain legally enforceable rights.

The same act can be and very often is both negligent and a breach of contract. In fact, negligence in the nature of malpractice alleged by a client against an engineering firm will almost invariably constitute a breach of contract as well as negligence, since the engineer, by contracting with the client, undertakes to comply with the standard of practice employed by average local engineers. If the condition is not expressed, it is generally implied by the courts.

Insurance for Engineers

It is customary for most businesses and some individual engineers to carry comprehensive liability insurance. The insurance industry recognizes that engineers, because of their occupation, are susceptible to special risks of liability. Therefore, when a carrier issues a comprehensive liability policy to an engineering consultant or firm, it may exclude from the insurance afforded by the policy the risk of professional negligence, malpractice, and “errors and omissions.” The engineer should seek independent advice on the extent and type of the coverage being offered before accepting coverage. However, depending on the wording of the policy and the specific nature of the claim, the comprehensive liability carrier may be under a duty to defend an action against the insured and sometimes must also pay the loss. When a claim is made against an insured engineering consultant or firm, the consultant or firm should retain a competent attorney to review the policy prior to accepting the conclusions of the insurance agent as to the absence of coverage.

While the engineer employee of a well-insured firm probably has limited liability exposure, the professional engineering consultant should be covered by professional liability (malpractice) insurance. However, many engineers decide to forego malpractice insurance because of high premium rates. Claims may be infrequent but can be economically devastating when incurred. The proper amount of coverage is something that should be worked out with a competent underwriter and will vary by engineering discipline and type of work. A policy should be chosen that not only pays damages but also underwrites the costs of attorney’s fees, expert witnesses, and so on.

Case Study*

The following case serves to illustrate the importance of developing a fundamental understanding of the professional liability concepts discussed above.

S&W Engineering was retained by Chesapeake Paper Products to provide engineering services in connection with the expansion of Chesapeake's paper mill. S&W's vice president met with Chesapeake's project manager and provided him with a proposed engineering contract and price quotations. Several weeks later, Chesapeake's project manager verbally authorized S&W to proceed with the work. S&W's engineering contract was never signed by Chesapeake; instead, Chesapeake sent S&W a purchase order (PO) that authorized engineering services "in accordance with the terms and conditions" of S&W's engineering contract. However, Chesapeake's PO also contained language in smaller print stating "This order may be accepted only upon the terms and conditions specified above and on the reverse side."

The drawings supplied by S&W to Chesapeake's general contractor subsequently contained errors and omissions, resulting in delays and increased costs to Chesapeake. Chesapeake sued S&W for breach of contract, arguing that the PO issued by Chesapeake constituted the parties' contract and that this PO contained a clause requiring S&W's standard of care to be "free from defects in workmanship." Additionally, another PO clause required indemnification of all expenses "which might incur as a result of the agreement."

S&W agreed that its engineering drawings had contained some inconsistencies but denied that those errors constituted a breach of contract. S&W claimed that the parties' contract consisted of the terms in its proposed engineering contract it had delivered to Chesapeake at the outset of the project. S&W's engineering contract provided that the "Engineer shall provide detailed engineering services ... conforming with good engineering practice." S&W's proposed contract also contained a clause precluding the recovery of any consequential damages.

At a jury trial, 14 witnesses testified, and the parties introduced more than 1000 exhibits. The jury found that the parties' "operative contract" was the PO and that S&W's services did not meet the contractually required standard of care. Chesapeake was awarded \$4,665,642 in damages.

3 LAWS OF PRODUCT LIABILITY

3.1 Definition

In Section 1, the art of engineering was characterized as a progression from real-world product specifications to the world modified by assumptions. This assumed world permits establishing precise component design parameters. Finally, the engineer must attempt to return to the real world by using a "safety factor" to bridge the gap between the ideal, but artificial, world of precise design calculations to the real world of uncertainties in who, how, and where the product will actually function.

The laws of product liability sharpen and intensify this focus on product behavior in the real world. *Product liability* is the descriptive term for a legal action brought by an injured person (the plaintiff) against another party (the defendant) alleging that a product sold (or manufactured or assembled) by the defendant was in a substandard condition and that this substandard condition was a principal factor in causing the harm of the plaintiff.

The key term for the engineer is *substandard condition*. In legal parlance, this means that the product is alleged to contain a *defect*. During litigation, the product is put on trial so that the jury can decide whether the product contained a defect and, if so, whether the defect caused the injury.

* *Chesapeake Paper Products v. Stone & Webster Engineering*, No. 94-1617 (4th Cir. 1995).

The laws of product liability take a retrospective look at the product and how it functioned as it interacted with the persons who used it within the environment surrounding the product and the persons. Three legal principles generally govern the considerations brought to this retrospective look at the engineer's art:

1. Negligence
2. Strict liability
3. Express warranty and misrepresentation

3.2 Negligence

This principal is based on the conduct or fault of the parties, as discussed in Section 2.1. From the plaintiff's point of view, it asks two things: first, whether the defendant acted as a *reasonable person* (or company) in producing and selling the product in the condition in which it was sold and, second, if not, was the condition of the product a substantial factor in causing the plaintiff's injury.

The test of *reasonableness* is to ask what risks the defendant (i.e., designer, manufacturer, assembler, or seller) foresaw as reasonably occurring when the product was used by the expected population of users within the actual environment of use. Obviously, the plaintiff argues that if the defendant had acted reasonably, the product designer would have foreseen the risk actually faced by the plaintiff and would have eliminated it during the design phase and before the product was marketed. That is, the argument is that the defendant, in ignoring or not accounting for this risk in the design of the product, did not properly balance the risks to product users against the utility of the product to society.

It is the *reasonableness*, or lack thereof, of *the defendant's behavior* (in designing, manufacturing, or marketing the product or in communicating to the user through instructions and warnings) that is the question under the principle of negligence. These issues are fully discussed in Section 5.

3.3 Strict Liability

In contrast to negligence, strict liability ignores the defendant's behavior. It is, at least in theory, of no consequence whether the manufacturer behaved reasonably in designing, manufacturing, and marketing the product. The only concern here is the quality of the product as it actually functions in society.

Essentially, the question to be resolved by the jury under strict liability is whether the risks associated with the real-world use of the product by the expected user population exceed the utility of the product and, if so, whether there was a reasonable alternative to the design that would have reduced the risks without seriously impairing the product's utility or making it unduly expensive. If the jury decides that the risks outweighed the product's utility and a reasonable alternative to reducing the risk existed, then the product is judged to be in a *defective condition unreasonably dangerous*.

Under strict liability, a product is defective when it contains *unreasonable* dangers, and only unreasonable dangers in the product can trigger liability. While it is unlikely the marketing department will ever use the term in a promotion campaign, a product may contain *reasonable* dangers without liability. In the eyes of the law, a product whose only dangers are reasonable ones is *not* defective.

Stated positively, a product that does not contain unreasonable dangers is "reasonably safe"—and that is all the law requires. This means that any residual risks associated with the product have been transferred *appropriately* to the ultimate user of the product.

Section 5 discusses the methodology for uncovering unreasonable dangers associated with products.

3.4 Express Warranty and Misrepresentation

The third basic legal principle governing possible liability has nothing to do with either the manufacturer's conduct (negligence) or the quality of the product (strict liability). Express warranty and misrepresentation are concerned only with what is communicated to the potential buyer that becomes part of the "basis of the bargain."

An express warranty is created whenever any type of communication to the potential buyer describes some type of *objectively measurable* characteristic of the product. The following are some sample express warranties:

- This truck will last 10 years.
- This glass is shatterproof.
- This automatic grinder will produce 10,000 blades per hour.
- This transmission tower will withstand the maximum wind velocities and ice loads in your area.

If such a communication is, first, at least a part of the reason that the product was purchased and then, if reasonably foreseeable circumstances ultimately prove the communication invalid, there has been misrepresentation and the buyer is entitled to recover damages consistent with the failed promise. It does not matter one whit if the product cannot possibly live up to the promise. This is not the issue. It is the failure to keep a promise that becomes part of the basis of the bargain and that the buyer did not have sufficient expertise for not believing the promise that can trigger the liability.

Someone with a legal bent might argue, against the misrepresentation claim, that the back of the sales form clearly and unequivocally disclaims all liability arising from any warranties not contained in the sales document (i.e., the contract). The courts, when confronted with what appears to be a conflict between the express warranty communicated to the buyer and the fine print on the back of the document disclaiming everything, inevitably side with the buyer who believed the express warranty to the extent that it became a part of the "basis of the bargain."

The communications creating the express warranty can be in any form: verbal, written, visual, or any combination of these. In the old days, courts used to view advertising as mere puffing and rarely sided with the buyer arguing about exaggerated claims made about the product. In recent years, however, the courts have acknowledged that buying is engendered in large part by media representations. Now, when such representations can be readily construed as express warranties, the buyer's claim is likely to be upheld. It should also be noted that misrepresentation claims have been upheld when both the plaintiff and the defendant are sophisticated, the defendants have staffs of engineers and lawyers, and the dealings between the parties are characterized as "arm's length."

In precarious economic times, the exuberance of salespersons, in their quest to make the sale, may oversell the product and create express warranties that the engineer cannot meet. This can then trigger liability, despite the engineer's best efforts.

Because it is so easy to create, albeit unintentionally, an express warranty, all departments that deal in any way with a product must recognize this potential problem and structure methods and procedures to minimize its occurrence. This means that engineering, manufacturing, sales, marketing, customer service, and upper management must create a climate in which there is agreement among the appropriate entities that what is being promised to the buyer can actually be delivered.

4 NATURE OF PRODUCT DEFECTS

The law recognizes four areas that can create a “defective condition unreasonably dangerous to the user or consumer”:

1. Production or manufacturing
2. Design
3. Instructions
4. Warnings

4.1 Production or Manufacturing Flaws

A production or manufacturing defect can arise when the product fails to emerge as the manufacturer intended. The totalities of the specifications, tolerances, and so on, define the product and all of the bits and pieces that make it up, and collectively they prescribe the manufacturer’s intent for exactly how the product is to emerge from the production line.

If there is a deviation from any of these defining characteristics of the product (e.g., specifications, tolerances), then there exists a production or manufacturing flaw. If this flaw or deviation can cause the product to fail or malfunction under reasonably foreseeable conditions of use in a way that can cause personal injury and/or property damage and these conditions are within the expected performance requirements for the product, then the product is defective.

What is important to note here is that the deviation from the specifications must be *serious* enough to be able to precipitate the failure or malfunction of the product within the foreseeable uses and performance envelope of the product. To illustrate, we return to the crane described in the first section of this chapter.

Example 5 Truck–Crane—Flaw or Defect?

Suppose that a critical weld is specified to be 4 in. in length and to have full penetration. After a failure, the crane is examined and the weld is full penetration but only 3½ in. long, which escaped the quality inspectors. There is a deviation or flaw. However, whether this flaw rises to the level of defect depends on several considerations: First, what safety factor considerations entered into the design of the weld? It may be that the designer calculated the necessary weld length to be 3 in. and specified 4 in. to account for the uncertainties described in Section 1. Next, if it can be shown by the crane manufacturer that a 3½-in. weld was adequate for all reasonably foreseeable use conditions of the crane, then it could be argued that the failure was due to crane misuse and not due to the manufacturing flaw.

Alternatively, the plaintiff could argue that the engineer’s assumptions as to the magnitude of the safety factor did not realistically assess the uncertainty of the weight loads to be lifted; if they had done so, the minimum acceptable length would have been the 4 in. actually specified.

While this is a hypothetical example, it illustrates the interplay of several important elements that must be considered when deciding if a production flaw can rise to the level of a defect. Foreseeable uses and misuses of the product and the prescribed or implicit performance requirements are two of the most important.

4.2 Design Flaws

The standard for measuring the existence of a production flaw is simple. One needs only to compare the product’s attributes as it actually leaves the production line with what the manufacturer intended them to be by examining the manufacturer’s internal documents that prescribe the entire product.

To uncover a design flaw, however, requires comparing the correctly manufactured product with a standard that is not readily prescribed and is significantly more complex. The standard is a societal one in which the risks of the product are balanced against its utility to establish whether the product contains unreasonable dangers. If there are unreasonable dangers, then the design flaw becomes a defect.

In the crane example, assume that there has been a boom failure and that the crane met all of the manufacturer's specifications, that is, no manufacturing defect is alleged. The plaintiff alleges, instead, that if the boom had been fabricated from a heavier gauge as well as a stronger alloy steel, the collapse would have been avoided. The plaintiff's contention can be considered a design flaw. There is no question that the boom could have been fabricated using the plaintiff's specifications and, for the sake of our discussion, we will also assume the boom would not have failed using the different material.

The critical question, however, is should the boom have been designed that way? The answer is, only if the original design created unreasonable dangers. The existence of unreasonable dangers, therefore a defective condition, can be deduced from a risk/utility analysis of the interaction of crane uses, users, and the environments within which the crane is expected to function.

The analysis must consider, first, the foreseeability of crane loads of uncertain magnitude that could cause the original design to fail, but not the modified design. Balanced against that consideration will be a reduction in the utility of the crane because of its increased weight and/or size if the proposed design alterations are incorporated. There will be also an increased cost. It is this analysis of competing trade-offs that the designer must consider before deciding on the proposed design specifications. Fundamentally, though, as in the discussion of a production defect, the consideration is that of the safety factor, bridging the gap between *assumed* product function and *actual* product function.

4.3 Instructions and Warnings

A product can be perfectly manufactured from a design that contains no unreasonable dangers and yet is defective because of inadequate instructions or warnings. Instructions are the communications between the manufacturer and the user that describe how the product is to be used to achieve the intended function of the product.

Warnings are to communicate any residual hazards, the consequent risks of injury, and the actions the user must take to avoid injury. If the warnings are inadequate, the product can be defective even if the design, manufacturing, and instructions meet the legal tests.

While the courts have not given clear or unequivocal guidelines for assessing the adequacy of instructions and warnings, there are several basic considerations that are drawn from litigated cases that must underlie their development to meet the test of adequacy:

- They must be understood by the expected user population.
- They must be effective in a multilingual population.
- There must be some reasonable and objective evidence to prove that the warnings and instructions can be understood and are likely to be effective.

Simply put, writing instructions and warnings is easy. However, gathering evidence to support the contention that they are *adequate* can be extremely difficult, costly, and time consuming.

To do this means surveying the actual user population and describing those characteristics that are likely to govern comprehension, such as age, education, reading capability, sex, and cultural and ethnic background. Then a statistically selected, random sample of the identified user population must be chosen to test the communication for comprehension using the method

suggested in the American National Standards Institute (ANSI) standard Z535.3. Finally, the whole process must be documented. Then, and only then, can a manufacturer argue that the user communications, that is, instructions and warnings, are adequate.

5 UNCOVERING PRODUCT DEFECTS

5.1 Hazard Analysis

In the preceding section, a risk/utility analysis was described as a basis for assessing whether or not the product was in a defective condition unreasonably dangerous. Consider now the methodology and the process of the risk/utility analysis.

We begin with a disclaimer: Neither the process nor the methodology about to be discussed is readily quantifiable. However, this fact does not lessen their importance; it only emphasizes the care that must be exercised.

The process is one of scenario building. The first step is to characterize, as accurately as possible, the users of the product, the ways in which they will use the product, and the environment in which they will use it. These elements must be quantified as much as possible.

Example 6 Foreseeable Users of a Hand-Held Tool.

Will the user population comprise younger users, female users, elderly users? If so, these populations are likely to need special ergonomic or human factors considerations in the design of handgrips, operating controls, etc. Will the tool be found in the home? If so, inadvertent use by small children is likely to be a consideration in designing the controls. Certainly the ability to read and understand instructions and warning must be a significant element of the characterization of the users.

In the best of all worlds, the only product uses the engineer would be concerned with are the *intended* uses. Unfortunately, the law requires that the product design acknowledge and account for *reasonably foreseeable misuses* of the product. Of all the concepts the engineer must deal with, this one is perhaps the hardest to analyze and the most difficult to accept. Part of the reason, of course, is the difficulty of distinguishing between uses that are *reasonably foreseeable* and those uses that the manufacturer can argue are truly *misuse* for which no account must be taken in design.

The concept of legal unforeseeability is a difficult one. Many people might think that if they have ever talked about the possibility of misusing a product in a certain way, then they have “foreseen” that misuse and therefore must account for it in their design. This is not the case. Legally, *unforeseeable misuse* means a use so egregious, or so bizarre, or so remote that it is termed *unforeseeable*, even when such a misuse has been a topic of discussion.

A simple illustration might help.

Example 7 How Many Ways Can You Use a Screwdriver?

There is no question that the intended purpose and function of a screwdriver is to insert and remove screws. This means that, ideally, the shank of a screwdriver is subjected only to a twisting motion, or torque.

But how do most people open paint cans? With a screwdriver, of course. In that context, however, the shank is subjected to a bending moment, not a torque. Any manufacturer who produced and marketed a screwdriver with shank material able to withstand high torque but without sufficient bending resistance to open a paint can without shattering would have a difficult time avoiding liability for any injuries that occurred.

The reason, of course, is that using a screwdriver to open paint cans would be considered as a reasonably foreseeable misuse and should be accounted for in the design. On the other hand,

suppose someone uses a screwdriver as a cold chisel to loosen a rusted nut and the screwdriver shatters, causing injury. The manufacturer could argue that such a use was a misuse for which the manufacturer had no duty to account for in the design.

Finding the line that separates the misuses the engineer must account for from the misuses that are legally unforeseeable is not easy, nor is the line a precise one. All that is required, however, is for the engineer to show the reasonableness of the process of how the line was ultimately decided, while attempting to meet competing trade-offs in selecting the product's specifications. Unquestionably, we can always imagine all types of bizarre situations in which a product is misused and someone is injured. Does this mean that all such situations must somehow be accounted for in design? Of course not. But what is required is to make a reasonable attempt to separate user behavior into two categories: that which can reasonably be accounted for in design and that which is beyond reasonable considerations.

The third element in the risk/utility process is the environment within which the user and product interact. If it is cold, how cold? If it is hot, how hot? Will it be dark, making warnings and instructions difficult to read? Will the product be used near water? If so, both fresh and salt? How long will the product last? Will it be repainted, scraped, worn, and so on? These, too, would be considerations in warning adequately.

The scenario building must integrate the three elements of the hazard analysis: the users, the uses, and the environment. By asking "What if ... ?" a series of hazards can be postulated from integrating the users with the uses within an environment.

Example 8 "What If an Air-operated Sander ... ?"

What if an air-operated sander is used in a marine environment? What if the user inadvertently drops it overboard and then continues to use it without having it disassembled and cleaned? What hazards could arise? Could corrosion ultimately freeze the control valve continually open, leading to loss of control at some future time, long after the event in question?

5.2 Hazard Index

After completing the hazard analyses, the hazards should be rank ordered from the most serious to the least serious. One way to do this is to assign a numerical probability of the event occurring and then to assess, also using a numerical scale, the seriousness of the harm. The product of these two numbers is the *hazard index* and permits a relative ranking of the hazards. The scales chosen to provide some measure of probability and seriousness should be limited; the scale may run, for example, from 0 to 4. A 0 implies that the event is so unlikely to occur or the resulting harm so minimal as to be negligible. Correspondingly, a 4 would mean that an event was almost certain to occur or that the result would be death or serious irreparable injury. With this scale, the hazard index could range from 0 to 16.

Once this is done, attention is then focused on the most serious hazards, eventually working down to the least serious one.

5.3 Design Hierarchy

Ideally, for each such event, the objective would be, first, to "design out" the hazard. If a hazard can be designed out, it can never return to cause harm.

Failing the ability of designing out the hazard, the next consideration must be guarding. Can an unobtrusive barrier be placed between the user and the hazard? It must be noted that if a guarding configuration greatly impairs the utility of the product or greatly increases the time needed to carry out the product's intended function, it is likely to be removed. In such a case, the user is not protected from the hazard, nor is the manufacturer likely to be protected

from liability if an injury results, because removing an obtrusive guard may be considered a foreseeable misuse.

If the hazard cannot be designed out or an effective guard cannot be devised, then *and only then* should the last element of the design hierarchy be considered: a warning. A warning must be viewed as an element of the design process, not as an afterthought. To be perfectly candid, if the engineer has to resort to a warning to minimize or eliminate a risk of injury from that hazard, it may be an admission of a failure in the design process.

Yet, there are innumerable instances where a warning must be given. Section 4 described the considerations necessary to develop an adequate warning, the legal standard. What was not described there are the three necessary elements that must be included before the process of establishing adequacy begins:

1. Nature of the hazard
2. Potential magnitude of the injury
3. Action to be taken to avoid the injury

A warning paraphrased from an aerosol can of hair spray provides an exercise for the reader:

Warning

- Harmful vapors
- Inhalation may cause death or blindness
- Use in a well-ventilated area

The reader should analyze these three warnings carefully and critically, then describe the user populations to which they might apply. Then the question of whether or not it is likely that injury could be avoided by that user population needs to be answered. Suppose that a foreseeable portion of the population using this aerosol can are people whose English reading ability is at the third or fourth grade level. (It is estimated that about half of the English-speaking Americans cannot read beyond the fourth grade level.) What can you conclude about comprehension and the ability to avoid injury?

Warnings are, in fact, the most difficult way to minimize or eliminate hazards to users.

6 DEFENSE TO PRODUCT LIABILITY

Up to now, we have looked only at the factors that permit an analysis of whether the product contains a defect, i.e., an unreasonable danger. Certainly the ultimate defense to an allegation that the product was defective is to show through a risk/utility analysis that, on balance, the product's utility outweighs its risks and, in addition, there were no feasible alternatives to the present design. It may be, however, that the plaintiff's suggested design alternative is, in fact, viable as of the time the incident occurred. Is there any analysis that could offer a defense? There may be, by considering a *state-of-the-art* argument.

6.1 State of the Art

Decades ago, the term *state of the art* meant, simply, what the custom and practice was of the particular industry in question. Because of the concern that an entire industry could delay introduction of newer, safer designs by relying on the "custom and practice" argument to defeat a claim of negligence, the courts have adopted a broader definition of the term. The definition today is "what is both technologically and economically feasible." The time at which this analysis is performed is, in general, the date the product in question was manufactured. Thus, while

a plaintiff's suggested alternative design may have been technologically and economically feasible at the time the incident occurred, their argument may not be viable if the product was manufactured 10 years before the incident occurred.

To make that argument convincing, however, means that engineers must always be actively seeking new and emerging technology, looking to its potential applicability to their industry and products. It is expected, too, that technological advances are sought, not only in the engineer's own industry, but in related and allied fields as well. Keeping current has an added dimension, that of being alert to broader vistas of technological change outside one's own industry.

The second element of today's state-of-the-art principle is that innovative advances must be economically viable as well. It is generally, but incorrectly, assumed that the term *economic viability* is limited to the incremental cost of incorporating the technological advance into the product and how it will affect the direct cost of manufacturing and the subsequent profit margin. The courts, however, are concerned with another cost in measuring economic viability, in addition to the direct of incorporating a safety improvement in the product: the cost to society and ultimately to the manufacturer if the technological advance is *not* incorporated into the product and injuries occur as a result. The technological advances we are concerned with here are those that are likely to enhance safety.

While it is more difficult and certainly cannot be predicted with a great deal of precision, an estimate of costs of the probable harm to product users is part of the equation. An approach to this analysis was described in Section 5. Estimating both the probability and seriousness of the harm from a realistic vantage point if the technological advance is *not* incorporated can form the basis for estimating the downside risk of not including the design feature.

6.2 Role of Safety Standards

Very often, the plaintiff's expert can argue, and with credibility we will assume that a design modification or additional guard would have prevented the plaintiff's injury and that such a change would have been both technologically and economically feasible at the time of the manufacture. If so, the manufacturer is unlikely to prevail with a state-of-the-art argument as a defense. Is there somewhere else a manufacturer can turn for a possible defense?

Few products that are manufactured today do not conform, at least in part, to some type of safety standards. There are a variety of these standards that often play a role in how the product is designed and manufactured. Among these are U.S. government standards such as those mandated by the Department of Transportation, Coast Guard, Food and Drug Administration, and so on. In addition, there are voluntary and consensus standards such as those promulgated by Underwriters Laboratories (UL), the ANSI, and the American Society for Testing and Materials (ASTM). Finally, there are standards developed by and for a given industry such as the Food Equipment Manufacturers Association or the Conveyor Manufacturers Association.

Assume that a product for which the plaintiff's expert has proposed a modification conforms to all applicable safety standards and that the standards do not require the plaintiff's expert's suggested modification. Then, can a manufacturer prevail by arguing adherence to those standards? With nothing more than the absence of a requirement in the standard to support the manufacturer's claim of a nondefective product, the standard is unlikely to shield the manufacturer from liability. The reason, quite simply, is that courts generally view standards as floors, not ceilings. That is, conforming to standards, even those promulgated by the federal government, is considered by the courts to be the *least* a manufacturer should do to produce a safe product. The court in a case involving a backyard trampoline (*Dudley Sports Co. v. Schmitt*) stated:

The fact that a particular product meets or exceeds the requirements of its industry is not conclusive proof that the product is reasonably safe. In fact, standards set by an entire industry can be found negligently low if they fail to meet the test of reasonableness.

This is the view of standards generally held by the courts. The fact that federal standards are subject to the same threshold-of-safety consideration reflects the view that all standards, including federal ones, are developed by a process dominated by the industry that is expected to conform to the standards. As a result, the standards emerging from the process are likely to reflect a limited consideration of hazard identification and risk assessment in deciding on the extent of safety requirements.

If a manufacturer is to overcome the stigma that the product standard is unlikely to reflect the requirements necessary for a reasonably safe product, it is necessary for the manufacturer first to demonstrate that it undertook an independent hazard identification and risk assessment for the product in question during the design process. Next, it must be shown that no unreasonably dangerous conditions emerged from the analysis that were not adequately covered by the standard's requirements. Finally, the manufacturer must present plausible arguments refuting the plaintiff's expert's proposals for a design or guarding modification that would have prevented the plaintiff's injury. Needless to say, such arguments must be substantive and avoid dismissing the proposed changes as being costly.

While standards can be a useful guide for the product design team, they must not be viewed as ends in themselves. Standards have rarely, if ever, shielded a manufacturer from liability absent an independent verification by the manufacturer that the standards defined a reasonably safe product.

6.3 Contributory/Comparative Negligence

We have not yet really considered what role, if any, the plaintiff's behavior plays in defending a product against an allegation of defect. We have earlier touched on misuse of the product, which is a use so egregious and so bizarre or so remote that it is termed *legally unforeseeable*. You may recall the example discussing the hypothetical use of a screwdriver as a cold chisel to illustrate what could very likely be considered as misuse.

But what about the plaintiff's behavior that is not so extreme? Does that enter at all into the equation of how fault is apportioned? Yes, it does, in the form of contributory or comparative negligence if the legal theory embracing the litigation is negligence. You will recall that under negligence the defendant's behavior is measured by asking if that party was acting as a *reasonable* person (or manufacturer or engineer) would have acted under the same or similar circumstances. And the reasonableness of the behavior is the result of having foreseen the risks of one's actions by having undertaken a risk/utility balancing prior to engaging in the action.

In a negligence action, the plaintiff's behavior is measured in exactly the same way. The defendant asks the jury to consider whether the plaintiff was behaving as a reasonable person would have under the same or similar circumstances. Did the plaintiff contribute to his or her harm by not acting reasonably? This is called *contributory negligence*.

While some states still retain the original concept that *any* contributory negligence on the part of the plaintiff totally bars his or her recovery of damages, most states have adopted some form of comparative negligence. Generally, the jury is asked to assess the behavior of both the plaintiff and the defendant and apportion the fault in causing the harm between them, making certain the percentages total 100%. The plaintiff's award, if any, is then reduced by the percentage of his or her comparative negligence.

The test of the defendant's negligence and the plaintiff's contributory negligence is termed an objective one. That is, the jury is asked to judge the actions of the parties relative to what a reasonable person would have done in the same or similar circumstances. The jury does not, as a rule, consider whether anything in that party's background, training, age, experience, education, and so on, played any role in the actions that led to the injury.

6.4 Assumption of Risk

There is another defense involving the plaintiff's behavior that does consider the plaintiff's characteristics in assessing his or her culpability. It is termed *assumption of the risk*. In essence, this defense argues that the plaintiff consented to being injured by the product. In one common form, used for analyzing this aspect of the plaintiff's behavior, the jury is asked if the plaintiff *voluntarily* and *unreasonably* assumed a *known* risk. To prevail, the defendant must present evidence on all three of these elements and must prevail on all three for a jury to conclude that the plaintiff "assumed the risk."

The first element, asking whether the plaintiff voluntarily confronted the danger, and the third element, considering whether the risk was known, are both subjective elements. That is, the jury must determine the state of the mind of the plaintiff, assessing what he or she actually knew or believed or what can reasonably be inferred about his or her behavior at the instant prior to the event that led to injury. Thus, the background, education, training, experience, and so on, become critical elements in this assessment.

A couple of points should be made here. First, in determining whether the plaintiff voluntarily confronted the hazard, the test is whether or not the plaintiff had *viable* alternatives.

Example 9 Work or Walk.

In a workplace setting, a worker is given a choice of either using a now-unguarded press or being fired. It had been properly guarded for all the time the plaintiff had used it in the past, but the employer had removed the guards to increase productivity and now tells the employee either to use the press as is or be fired. The courts do not consider that the plaintiff had viable alternatives, since the choice between working on an unguarded press or being fired is no choice at all. The lesson to the engineer in this example is that the guarding slowed productivity and was removed, leaving the press user in a no-win situation. The design should have incorporated guarding that did not *slow production*.

Second, the same in-depth consideration must also be given to knowledge of the risk by the plaintiff. The plaintiff's background, education, and so, on must provide a reasonable appreciation of the actual nature of the harm that could befall him or her.

Example 10 The Truly Combustible Car.

The driver of a new car is confronted by a slight smell of smoke the first time the windshield wipers are used and is trying to bring the car to the dealer to see what the trouble is when the car bursts into flames, causing injury. Has the driver assumed the risk of injury by continuing to drive after smelling smoke? Can the car manufacturer successfully argue that the risks of injury were known to the driver? The question can be answered only by examining those elements in the driver's background that could, in any way, lead a jury to conclude that the driver should have recognized that smoke from electrically operated wipers could lead to a conflagration. The old adage of "where there's smoke, there's fire" is insufficient to charge the plaintiff with knowledge of the precise risk he or she faced without more knowledge of the driver's background.

The final element of assumption of the risk, the unreasonableness of the plaintiff's choice in voluntarily confronting a known risk, is an objective element, exactly the same as in negligence. That is, what would a reasonable person have done under the same or similar circumstances?

Example 11 The Truly Combustible Car Meets the Good Samaritan.

A passer-by observes the car from the previous example. It is on fire, and the driver is struggling to get out. The passer-by rescues the driver but is seriously burned and suffers smoke inhalation in the process. The driver files suit against the manufacturer alleging a defect that created

unreasonable danger when the wipers were turned on. The passer-by also files suit against the automobile manufacturer to recover for the injuries suffered as a result of the rescue, arguing that the rescue would not have been necessary if there had been no defect. Would this good Samaritan be found to have assumed the risk of injury? Clearly the choice to try to rescue the driver was voluntary and the risks of injury from a fire were apparent to anyone, including the rescuer. But was the act of rescuing the car's occupant a reasonable or unreasonable one? If the jury concludes that it was a reasonable choice, the passer-by would not have been found to have assumed the risk, despite having voluntarily exposed himself to a known risk.

The defendant must prevail in all three of the elements, not just two. Needless to say, raising and succeeding in the defense of assumption of the risk are not easy for the defendant.

One final word about these defenses: While the "assumption of the risk" defense applies both in a claim of negligence and strict liability, the contributory/comparative negligence defense does *not* apply in strict liability. The reason is that strict liability is a no-fault concept, whereas negligence is a fault-based concept. It would be inconsistent to argue no-fault theory (strict liability) against the defendant and permit the defendant to argue a fault-based defense (contributory negligence) concerning the plaintiff's behavior.

7 RECALLS, RETROFITS, AND CONTINUING DUTY TO WARN

Manufacturers generally have a postsale or continuing duty to warn of latent defects in their products that are revealed through consumer use. Sometimes, however, even a postsale warning may be inadequate to render a product safe. In those circumstances, it may be necessary for a manufacturer to retrofit the product by adding certain safety devices or guards. Moreover, there may be instances where it is not feasible to add guards or safety devices or where the danger of the product is so great that the product simply must be removed from the market by being recalled.

7.1 After-Market Hazard Recognition

The manufacturer is responsible for establishing feedback mechanisms from customers, distributors, and sales personnel that will ensure that postsale problems are discovered. Applicable data may include product performance and test data, orders for repair parts, complaint files, quality control and inspection records, and instruction and warning modifications. Another source of hazard recognition information comes from previous accident investigations, claims, and lawsuits. The manufacturer should also have an ongoing program of compiling and evaluating risk data from historical, field and/or laboratory testing, and fault tree, failure modes, and hazard analyses.

Once the manufacturer has determined that a previously sold product is defective (that is, contains unreasonable dangers) and still in use, it must decide on what response is appropriate. If the product is currently being produced, an initial assessment as to the seriousness of the problem must be made to decide whether production is to be halted immediately and inventories frozen in the warehouses and on dealers' shelves in order to limit distribution.

Following this assessment, the nature of the defect must be established. If the problem is safety related and depending on the type of the product, appropriate regulatory agencies may have to be immediately notified. The manufacturer must then consider the magnitude of the hazards by estimating the probability of occurrence of events and the likely seriousness of injury or damage. The necessity for postulating such data is to provide some measure of the magnitude of the consequences if no action is taken or to decide the extent of the action to be taken in light of the estimated consequences. Alternatively, if the consequences of even a low probability of occurrence could result in serious injury or death or could seriously affect the

marketability of the product or the corporate reputation, the decision to take action should be independent of such estimates.

Once the decision to take action is made, the origin, extent, and cause of the problem must be addressed in order to plan effective corrective measures. Is the origin of the defect in the raw material, fabrication, or quality control? If the problem is one of fabrication, did it occur in-house or from a purchased part? Where are the faulty products—that is, are the products in inventory, in shipment, in dealers' stock, or in the hands of the buyers? Does the defect arise from poor design, inadequate inspection, improper materials, fabrication procedure, ineffective or absent testing, or a combination of these events?

7.2 Types of Corrective Action

After the decision to take action has been made and the origin, extent, and cause of the problem have been investigated, the appropriate corrective action must be determined. Possible options are to recall the product and replace it with another one; to develop a retrofit and either send personnel into the field to retrofit the product or have the customer return the product to the manufacturer for repair; to send out the parts and have the customer fix the product; or to simply send out a warning about the problem. This process should be fully documented to substantiate the reasons for the selection of a particular response. The urgency with which the corrective action is taken will be determined by the magnitude of the hazard.

Warnings

A manufacturer is not required to warn of every known danger, even with actual knowledge of that danger. A warning is required where a product can be dangerous for its intended and reasonably foreseeable uses and where the nature of the hazard is unlikely to be readily recognized by the expected user class. When a hazard associated with a product that was previously unknown to the manufacturer becomes apparent after the product has been in use, the manufacturer has a threshold duty to warn the existing user population.

Factors to consider in determining whether to issue a postsale warning include the manufacturer's ability to warn (i.e., how readily and completely the product users can be identified and located), the product's life expectancy (the longer the life expectancy, the greater the risk of potential harm if postsale warnings are not given), and the obviousness of the danger. Thus, the practicality, cost, and burden of providing an effective warning must be weighed against the potential harm of omitting the warning.

Recalls

Where the potential harm to the consumer is so great that a warning alone is not adequate to eliminate the danger, the proper remedy may be to institute a recall of the product for either repair or replacement. For some products, a recall may be mandated by statute or a governmental regulatory agency. Where a recall is not mandated, however, the decision to institute a product recall should be made using the analysis undertaken in Section 7.1.

Retrofits

A recall campaign may not be an appropriate solution, particularly if the equipment is large or cannot be easily removed from an installation. For equipment with potentially serious hazards or requiring complicated modification, the manufacturer should send its personnel to perform (and document) the retrofit. For equipment with relatively minor potential hazards for which there is a simple fix, the manufacturer may opt to send to the owners the parts necessary to solve the problem.

Regardless of the type of corrective action program selected, it is essential that all communications directed to the owners and/or users urging them to participate in the corrective action program be clear and concise. Most importantly, however, is the necessity for the communication to identify the nature of the risks and the potential seriousness of the harm that could befall the product user.

8 DOCUMENTATION OF DESIGN PROCESS

There are conflicting arguments by attorneys about what documentation, if any, the manufacturer should retain in the files (or on the floppies, the discs, the hard drive, or tape backup). Since it would be well-nigh impossible to run a business without documentation of some sort, it only makes sense to preserve the type of documentation that can, if the product is challenged in court, demonstrate the care and concern that went into the design, manufacturing, marketing, and user communications of the product.

The first principle of documentation is to minimize or eliminate potential adverse use of the documentation by an adverse party. For example, words such as *defect* should not appear in the company's minutes, notes, and so on. There can be *deviations, flaws, departures*, and so on, from specifications or tolerances. These are not defects unless they could create unreasonable dangers in the use of the product. Also, all adverse criticism of the product, whether internally from employees or externally from customers, dealers, and so on, must be considered and addressed in writing by the responsible corporate person having the appropriate authority.

Apart from these considerations, the company should make an effort to create a *documentation tree*, delineating what document is needed, who should write it, where it should be kept, who should keep it, and for how long. The retention period for documents, for the most part, should be based on common sense. If a government or other agency requires or suggests the length of time certain documents be kept, obviously those rules must be followed. For the rest, the length of time should be based on sound business practices. If the product has certain critical components that, if they fail before the end of the product's useful life, could result in a serious safety problem, the documentation supporting the efficacy of these parts should be retained for as long as the product is likely to be in service.

Because the law requires only that a product be reasonably safe, clearly the documentation to be preserved should be that which will support the argument that all of the critical engineering decisions, which balanced competing trade-offs, were reasonable and were based on reducing the risks from all foreseeable hazards. The rationales underlying these decisions should be part of the record for two reasons. First, it will give those who will review the designs when the product is to be updated or modified in subsequent years the bases for existing design decisions. If the prior assumptions and rationales are still valid, they need not be altered. Conversely, if some do not reflect current thinking, then only those aspects of the design need to be altered. Without these rationales, all the design parameters will have to be reexamined for efficacy.

Second, and just as important, having the rationales in writing for those safety-critical decisions can provide a solid, legal defense if the design is ever challenged as defective.

Thus, the documentation categories that are appropriate for both subsequent design review and creating strong legal defense positions are these:

- Hazard and risk data that formed the bases for the safety considerations
- Design safety formulations, including fault tree and failure modes and effects analyses
- Warnings and instructions formulation, together with the methodology used for development and testing
- Standards used, including in-house standards, and the rationale for the requirements utilized in the design

- Quality assurance program, including the methodology and rationale for the processes and procedures
- Performance of the product in use, describing reporting procedures, follow-up data acquisition and analysis, and a written recall and retrofit policy

This type of documentation will permit re-creating the process by which the reasonably safe product was designed, manufactured, and marketed.

9 FINAL WORD

In the preceding pages, we have only touched on a few of the areas where the law can have a significant impact on engineers' discharge of their professional responsibilities. As part of the process of product design, the law asks the engineer to consider that the product which emerges from the mind of the designer and the hand of the worker to play a role in enhancing society's well-being must:

- Account for reasonably foreseeable product misuse
- Acknowledge human frailties and the characteristics of the actual users
- Function in the true environment of product use
- Eliminate or guard against the hazards
- Not substitute warnings for effective design and guards

What has been discussed here and summarized above is, after all, just good engineering. It is to help the engineer recognize those considerations that are necessary to bridge the gap between the preliminary product concept and the finished product that has to function in the real world, with real users and for real uses, for all of its useful life.

Apart from understanding and utilizing these considerations during the product design process, engineers have an obligation, both personally and professionally, to maintain competence in their chosen field so that there can be no question that all actions, decisions, and recommendations, in retrospect, were reasonable.

That is, after all, what the law requires of all of us.

CHAPTER 26

PATENTS

David A. Burge
David A. Burge Company
Cleveland, Ohio

Benjamin D. Burge
Intel Americas, Inc.
Chantilly, Virginia

1	WHAT DOES IT MEAN TO OBTAIN A PATENT?	774	3.9	Small-Entity Status	788
1.1	Utility, Design, and Plant Patents	774	3.10	Express Mail Filing	789
1.2	Patent Terms and Expiration	774	4	PROSECUTING PENDING PATENT APPLICATION	789
1.3	Four Types of Applications	775	4.1	Patent Pending	789
1.4	Why File Provisional Application?	775	4.2	Publication of Pending Applications	790
1.5	Understanding That a Patent Grants a “Negative Right”	776	4.3	Duty of Candor	791
2	WHAT CAN BE PATENTED AND BY WHOM	777	4.4	Initial Review of Application	791
2.1	Ideas, Inventions, and Patentable Inventions	777	4.5	Response to Office Action	792
2.2	Requirement of Statutory Subject Matter	778	4.6	Reconsideration in View of Filing of Response	793
2.3	Requirement of Originality of Inventorship	779	4.7	Interviewing Examiner	793
2.4	Requirement of Novelty	780	4.8	Restriction and Election Requirements	794
2.5	Requirement of Utility	782	4.9	Double-Patenting Rejections	794
2.6	Requirement of Nonobviousness	782	4.10	Continuation, Divisional, and Continuation-in-Part Applications	794
2.7	Statutory Bar Requirements	783	4.11	Maintaining Chain of Pending Applications	795
3	PREPARING TO APPLY FOR A PATENT	783	4.12	Patent Issuance	795
3.1	Patentability Search	784	4.13	Safeguarding Original Patent Document	796
3.2	Putting Invention in Proper Perspective	784	4.14	Reissue	796
3.3	Preparing Application	785	4.15	Reexamination	796
3.4	Enablement, Best Mode, Description, and Distinctness Requirements	785	5	ENFORCING PATENTS AGAINST INFRINGERS	797
3.5	Product-by-Process Claims	786	5.1	Patent Infringement	797
3.6	Claim Format	786	5.2	Defenses to Patent Enforcement	798
3.7	Executing Application	787	5.3	Outcome of Suit	798
3.8	U.S. Patent and Trademark Office Fees	788	5.4	Settling Suit	799
			5.5	Declaratory Judgment Actions	799
			5.6	Failure to Sue Infringers	799

5.7	Infringement by Government	800	6.3	Annual Maintenance Taxes and Working Requirements	801
5.8	Alternative Resolution of Patent Disputes	800	6.4	Filing under International Convention	802
5.9	Interferences	800	6.5	Filing on Country-by-Country Basis	802
6	PATENT PROTECTIONS AVAILABLE ABROAD	801	6.6	Patent Cooperation Treaty	802
6.1	Canadian Filing	801	6.7	European Patent Convention	803
6.2	Foreign Filing in Other Countries	801	6.8	Advantages and Disadvantages of International Filing	803

1 WHAT DOES IT MEAN TO OBTAIN A PATENT?

Before meaningfully discussing such topics as inventions that qualify for patent protection and procedures that are involved in obtaining a patent, it is necessary to know about such basics as the four different types of patent applications that can be filed, the three different types of patents that can be obtained, and what rights are associated with the grant of a patent.

1.1 Utility, Design, and Plant Patents

When one speaks of obtaining a patent, it is ordinarily assumed that what is intended is a utility patent. Unless stated otherwise, the discussion of patents presented in this chapter applies only to U.S. patents and principally to utility patents.

Utility patents are granted to protect processes, machines, articles of manufacture, and compositions of matter that are new, useful, and nonobvious.

Design patents are granted to protect ornamental appearances of articles of manufacture—that is, shapes, configurations, ornamentation, and other appearance-defining characteristics that are new, nonobvious, and not dictated primarily by functional considerations.

Plant patents are granted to protect new varieties of plants that have been asexually reproduced with the exception of tuber-propagated plants and those found in an uncultivated state. New varieties of roses and shrubs often are protected by plant patents.

Both utility and design patents may be obtained on some inventions. A utility patent typically will have claims that define novel combinations of structural features, techniques of manufacture, and/or methods of use of a product. A design patent typically will cover outer configuration features that are not essential to the function of the product, but rather give the product an esthetically pleasing appearance. Utility patents can protect structural, functional, and operational features of an invention. Design patents can protect appearance features of a product.

Genetically engineered products may qualify for plant patent protection, for utility patent protection, and/or for other protections provided for by statute that differ from patents. Computer software and other computer-related products may qualify for design and/or utility patent protections. Methods of doing business may qualify for utility patent coverage. These are developing areas of intellectual property law.

1.2 Patent Terms and Expiration

Design patents that currently are in force have normal terms of 14 years, measured from their issue dates. Prior to a change of law that took effect during 1982, it was possible for design patent owners to elect shorter terms of $3\frac{1}{2}$ –7 years.

Plant and utility patents that expired prior to June 8, 1995, had a normal term of 17 years, measured from their issue dates. Plant and utility patents that (1) were in force on June 8, 1995, or (2) issue from applications that were filed prior to June 8, 1995, have normal terms that expire either 17 years, measured from their issue dates, or 20 years, measured from the filing date of the applications from which these patents issued, whichever is later. Plant and utility patents that issue from applications that were filed on or after June 8, 1995, have normal terms that expire 20 years from application filing, but these terms may be adjusted by a few days by the U.S. Patent and Trademark Office (USPTO) in view of delays that were encountered during application pendency (see Section 4.12).

The filing date from which the 20-year measurement of the term of a plant or utility patent is measured to calculate the normal expiration date of a utility patent is the earliest applicable filing date. If, for example, a patent issues from a continuation application, a divisional application, or a continuation-in-part application that claims the benefit of the filing date of an earlier filed “parent” application, the 20-year measurement is taken from the filing date of the parent application.

The normal term of a patent may be shortened due to a variety of circumstances. If, for example, a court of competent jurisdiction should declare that a patent is “invalid,” the normal term of the patent will have been brought to an early close. In some circumstances, a “terminal disclaimer” may have been filed by the owner of a patent to cause early termination. The filing of a terminal disclaimer is sometimes required by the USPTO during the examination of an application that is so closely subject matter related to an earlier filed application that there may be a danger that two patents having different expiration dates will issue covering substantially the same invention.

1.3 Four Types of Applications

Three types of patent applications are well known. A *utility* application is what one files to obtain a utility patent. A *design* application is what one files to obtain a design patent. A *plant* application is what one files to obtain a plant patent.

Starting June 8, 1995, it became possible to file a fourth type of patent application known as a *provisional* application as a precursor to the filing of a utility application. The filing of a provisional application will *not* result in the issuance of any kind of patent. In fact, no examination will be made of the merits of the invention described in a provisional application. Examination “on the merits” takes place only if a utility application is filed within one year that claims the benefit of the filing date of the provisional; and, if examination takes place, it centers on the content of the utility application, not on the content of any provisional applications that are referred to in the utility application.

The filing of a provisional application that adequately describes an invention will establish a filing date that can be relied on in a later-filed utility application relating to the same invention (1) if the utility application is filed within one year of the filing date of the provisional application and (2) if the utility application makes proper reference to the provisional application. While the filing date of a provisional application can be relied on to establish a reduction to practice of an invention, the filing date of a provisional application does *not* start the 20-year clock that determines the normal expiration date of a utility patent.

Absent the filing of a utility application within one year from the filing date of a provisional application, the USPTO will destroy the provisional application once it has been pending a full year.

1.4 Why File Provisional Application?

If a provisional application will not be examined and will not result in the issuance of any form of patent whatsoever, why would one want to file a provisional application? Actually, there are

several reasons why the filing of one or more provisional applications may be advantageous before a full-blown utility application is put on file.

If foreign filing rights are to be preserved, it often is necessary for a U.S. application to be filed before *any* public disclosure is made of an invention so that one or more foreign applications can be filed within one year of the filing date of the U.S. application, with the result that the foreign applications will be afforded the benefit of the filing date of the U.S. application (due to a treaty referred to as the *Paris Convention*), thereby ensuring that the foreign applications comply with “absolute novelty” requirements of foreign patent law.

If the one-year grace period provided by U.S. law (which permits applicants to file a U.S. application anytime within a full one-year period from the date of the first activity that starts the clock running on the one-year grace period) is about to expire, it may be desirable to file a provisional application rather than a utility application, because (1) preparing and filing a provisional application usually can be done less expensively, (2) preparing and filing a provisional application usually can be done more quickly inasmuch as it usually involves less effort (i.e., there is no need to include an abstract, claims, or a declaration or oath, which are required in a utility application), and (3) everything that one may want to include in a utility application may not have been discerned (i.e., development and testing of the invention may still be underway), and hence, it may be desirable to postpone for as much as a full year the drafting and filing of a utility application.

If development work is still underway when a first provisional application is filed and if the result of the development program brings additional invention improvements to light, it may be desirable (during the permitted period of one year between the filing of the first provisional application and the filing of a full-blown utility application) to file one or more additional provisional applications, all of which can be referred to and can have their filing dates relied upon when a utility application is filed within one year, measured from the filing date of the earliest filed provisional application.

Yet another reason why a provisional application may be filed is to preserve patent rights in situations where the value of an invention is questionable and the cost of filing a full-blown utility case is not justified, but action must be taken soon or the deadline to file for patent protection will not be met. In such situations, it may be desirable to file a first provisional as simply as possible and follow by filing additional provisional cases as the invention is developed and its importance comes to be better understood. So long as a full-blown utility application is filed within a year of the filing of the first provisional, this approach can assist in salvaging patent rights in inventions that might otherwise have been passed over for patent protection.

What the filing of a provisional application preserves for one year is the right to file a utility application. It does not preserve the right to file a design application; therefore, if design patent rights are to be preserved, it may necessary to file a design application at the same time that one files a provisional application.

1.5 Understanding That a Patent Grants a “Negative Right”

A patent is a grant by the federal government to the inventor or inventors of an invention of certain rights—rights that are “negative” in nature. Patents issue only upon application, and only after rigorous examination to ensure that the invention meets certain standards and qualifications, and that the application meets certain formalities.

It is surprisingly common to find that even those who hold several patents fail to properly understand the “negative” nature of the rights that are embodied in the grant of a patent. What the patent grants is the “negative right” to “exclude others” from making, using, or selling an invention that is covered by the patent. *Not* included in the grant of a patent is a “positive right” enabling the patent owner to actually make, use, or sell the invention. In fact, a patent owner may be precluded, by the existence of other patents, from making, using, and selling his or

her patented invention. Illustrating this often misunderstood concept is the following example referred to as *The Parable of the Chair*:

If inventor A invents a three-legged stool at an early time when such an invention is not known to others, A's invention may be viewed as being "basic" to the art of seats, probably will be held to be patentable, and the grant of a patent probably will have the practical effect of enabling A both (1) to prevent others from making, using, and selling three-legged stools and (2) to be the only entity who *can* legally make, use, and sell three-legged stools during the term of A's patent.

If, during the term of A's patent, inventor B improves upon A's stool by adding a fourth leg for stability and an upright back for enhanced support and comfort (whereby a chair is born), and if B obtains a patent on his chair invention, the grant of B's chair patent will enable B to prevent others from making, using, and selling four-legged back-carrying seats. However, B's patent will do nothing at all to permit B to make, use, or sell chairs—the problem being that a four-legged seat having a back *infringes* A's patent because each of B's chairs *includes* three legs that support a seat, which is what A can exclude others from making, using, or selling. To legally make, use, or sell chairs, B must obtain a license from A.

And if, during the terms of the patents granted to A and B, C invents and patents the improvement of providing curved rocking rails that connect with the leg bottoms, and arms that connect with the back and seat, thereby bringing into existence a rocking chair, C can exclude others during the term of his patent from making, using, or selling rocking chairs, but must obtain licenses from both A and B in order to make, use, or sell rocking chairs, for a rocking chair includes three legs and a seat, and includes four legs, a seat, and a back.

Invention improvements may represent very legitimate subject matter for the grant of a patent. However, patents that cover invention improvements may not give the owners of these patents any right at all to make, use, or sell their patented inventions unless licenses are obtained from those who obtained patents on inventions that are more basic in nature. Once the terms of the more basic patents expire, owners of improvement patents then may be able to practice and profit from their inventions on an exclusive basis during the remaining portions of the terms of their patents.

2 WHAT CAN BE PATENTED AND BY WHOM

For an invention to be patentable, it must meet several requirements set up to ensure that patents are not issued irresponsibly. Some of these standards are complex to understand and apply. Let us simplify and summarize the essence of these requirements.

2.1 Ideas, Inventions, and Patentable Inventions

Invention is a misleading term because it is used in so many different senses. In one, it refers to the act of inventing. In another, it refers to the product of the act of inventing. In still another, the term designates a patentable invention, the implication mistakenly being that if an invention is not patentable, it is not an invention.

In the context of modern patent law, invention is the conception of a novel, nonobvious, and useful contribution followed by its reduction to practice. Conception is the beginning of an invention; it is the creation in the mind of an inventor of a useful means for solving a particular problem. Reduction to practice can be either actual, as when an embodiment of the invention is tested to prove its successful operation under typical conditions of service, or constructive, as when a patent application is filed containing a complete description of the invention.

Ideas, per se, are not inventions and are not patentable. They are the tools of inventors, used in the development of inventions. Inventions are patentable only insofar as they meet certain

criteria established by law. For an invention to be protectable by the grant of a utility patent, it must satisfy the following conditions:

1. Fit within one of the statutorily recognized classes of patentable subject matter
2. Be the true and original product of the person or persons seeking to patent the invention as its inventor or inventors
3. Be new at the time of its invention by the person or persons seeking to patent it
4. Be useful in the sense of having some beneficial use in society
5. Be nonobvious to one of ordinary skill in the art to which the subject matter of the invention pertains at the time of its invention
6. Satisfy certain statutory bars that require the inventor to proceed with due diligence in pursuing efforts to file and prosecute a patent application

2.2 Requirement of Statutory Subject Matter

As stated in the Supreme Court decision of *Kewanee Oil v. Bicron Corp.*, 416 U.S. 470, 181 U.S.P.Q. 673 (1974), no utility patent is available for any discovery, however useful, novel, and nonobvious, unless it falls within one of the categories of patentable subject matter prescribed by Section 101 of Title 35 of the United States Code. Section 101 provides that:

Whoever invents or discovers a new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof may obtain a patent therefore, subject to the conditions and requirements of this title.

The effect of establishing a series of statutory classes of eligible subject matter has been to limit the pursuit of patent protection to the useful arts. Patents directed to processes, machines, articles of manufacture, and compositions of matter have come to be referred to as utility patents, inasmuch as these statutorily recognized classes encompass the useful arts.

Three of the four statutorily recognized classes of eligible subject matter may be thought of as products, namely machines, manufactures, and compositions of matter. *Machine* has been interpreted in a relatively broad manner to include a wide variety of mechanisms and mechanical elements. *Manufactures* is essentially a catch-all term covering products other than machines and compositions of matter. *Compositions of matter*, another broad term, embraces such elements as new molecules, chemical compounds, mixtures, alloys, and the like. *Machine and manufactures* now arguably includes software either stored on a piece of media or being executed on a computer. *Manufactures* and *compositions of matter* arguably include such genetically engineered life forms as are not products of nature. The fourth class, *processes*, relates to procedures leading to useful results.

Subject matter held to be ineligible for patent protection includes printed matter, products of nature, ideas, and scientific principles. Alleged inventions of perpetual motion machines are refused patents. A mixture of ingredients such as foods and medicines cannot be patented unless there is more to the mixture than the mere cumulative effect of its components. So-called patent medicines are seldom patented.

While no patent can be issued on an old product despite the fact that it has been found to be derivable through a new process, a new process for producing the product may well be patentable. That a product has been reduced to a purer state than was previously available in the prior art does not render the product patentable, but the process of purification may be patentable. A new use for an old product does not entitle one to obtain *product* patent protection but may entitle one to obtain *process* patent protection, assuming the process meets other statutory requirements. A newly discovered law of nature, regardless of its importance, is not entitled to patent protection.

While the requirement of statutory subject matter falls principally within the bounds of 35 U.S.C. 101, other laws also operate to restrict the patenting of certain types of subject matter. For example, several statutes have been passed by Congress affecting patent rights in subject matter relating to atomic energy, aeronautics, and space. Still another statute empowers the director of the USPTO to issue secrecy orders regarding patent applications disclosing inventions that might be detrimental to the national security of the United States.

The foreign filing of patent applications on inventions made in the United States is prohibited until a license has been granted by the director of the USPTO to permit foreign filing. This prohibition period enables the USPTO to review newly filed applications, locate any containing subject matter that may pose concerns to national security, and after consulting with other appropriate agencies of government, issue secrecy orders preventing the contents of these applications from being publicly disclosed. If a secrecy order issues, an inventor may be barred from filing applications abroad on penalty of imprisonment for up to two years or a \$10,000 fine or both. In the event that a patent application is withheld under a secrecy order, the patent owner has the right to recover compensation from the government for damage caused by the secrecy order and/or for any use the government may make of the invention.

Licenses permitting expedited foreign filing are almost always automatically granted by the USPTO at the time of issuing an official filing receipt, which advises the inventor of the filing date and serial number assigned to his or her application. Official filing receipts usually issue within a month of the date of filing and usually bear a statement attesting to the grant of a foreign filing license.

2.3 Requirement of Originality of Inventorship

Under U.S. patent law, only the true and original inventor or inventors may apply to obtain patent protection. If the inventor has derived an invention from any other source or person, he or she is not entitled to apply for or obtain a patent.

The laws of our country are strict regarding the naming of the proper inventor or joint inventors in a patent application. When one person acting alone conceives an invention, he or she is the sole inventor and he or she alone must be named as the inventor in a patent application filed on that invention. When a plurality of people contribute to the conception of an invention, these persons must be named as joint inventors if they have contributed to the inventive features that are claimed in a patent application filed on the invention.

Joint inventorship occurs when two or more persons collaborate in some fashion, with each contributing to conception. It is not necessary that exactly the same idea should have occurred to each of the collaborators at the same time. Section 116 of Title 35 of the United States Code includes the following provision:

Inventors may apply for a patent jointly even though (1) they did not physically work together or at the same time, (2) each did not make the same type or amount of contribution, or (3) each did not make a contribution to the subject matter of every claim of the patent.

Those who may have assisted the inventor or inventors by providing funds or materials for development or by building prototypes under the direction of the inventor or inventors are not deemed to be inventors unless they contributed to the conception of the invention. While inventors may have a contractual obligation to assign rights in an invention to their employers, this obligation, absent a contribution to conception, does not entitle a supervisor or an employer to be named as an inventor. When a substantial number of patentable features relating to a single overall development have occurred as the result of different combinations of sole inventors acting independently and/or joint inventors collaborating at different times, the patent law places a burden on the inventors to sort out “who invented what.” If one patent application covering the entire development and naming all of the inventors jointly is deemed inappropriate, then

patent protection on the overall development may need to be pursued in the form of a number of separate patent applications, each directed to such patentable aspects of the development as originated with a different inventor or group of inventors. In this respect, U.S. patent practice is unlike that of many foreign countries, where the company for whom all the inventors work is often permitted to file a single patent application in its own name covering the overall development.

Misjoinder of inventors occurs when a person who is not a joint inventor has been named as such in a patent application. *Nonjoinder of inventors* occurs when there has been a failure to include a person who should have been named as a joint inventor. *Misdesignation of inventorship* occurs when none of the true inventors are named in an application. Only in recent years has correction of a misdesignation been permitted. If a problem of misjoinder, nonjoinder, or misdesignation has arisen without deceptive intent, provisions of the patent law permit correction of the error as long as such is pursued with diligence following discovery.

2.4 Requirement of Novelty

Section 101 of Title 35 of the United States Code requires that a patentable invention be new. What is meant by *new* is defined in Sections 102(a), 102(e), and 102(g). Section 102(a) bars the issuance of a patent on an invention “known or used by others in this country, or patented or described in a printed publication in this or a foreign country, before the invention thereof by the applicant for patent.” Section 102(e) bars the issuance of a patent on an invention “described in a patent granted on an application for patent by another filed in the United States before the invention thereof by the applicant for patent, or in an international application by another.” Section 102(g) bars the issuance of a patent on an invention that “before the applicant’s invention thereof ... was made in this country by another who had not abandoned, suppressed or concealed it.”

These novelty requirements amount to negative rules of invention, the effect of which is to prohibit the issuance of a patent on an invention if the invention is not new. The novelty requirements of 35 U.S.C. 102 should not be confused with the statutory bar requirements of 35 U.S.C. 102, which are discussed in Section 2.7. A comparison of novelty and statutory bar requirements of 35 U.S.C. 102 is presented in Table 1. The statutory bar requirements are distinguishable from the novelty requirements in that they relate not to the newness of the invention, but to ways an inventor, who would otherwise have been able to apply for patent protection, has lost that right by tardiness.

To understand the novelty requirements of 35 U.S.C. 102, one must understand the concept of anticipation. A claimed invention is anticipated if a single prior art reference contains all the essential elements of the claimed invention. If teachings from more than one reference must be combined to show that the claimed combination of elements exists, there is no anticipation, and novelty exists. Combining references to render a claimed invention unpatentable brings into play the nonobviousness requirements of 35 U.S.C. 103, not the novelty requirement of 35 U.S.C. 102. Novelty hinges on anticipation and is a much easier concept to understand and apply than that of nonobviousness.

35 U.S.C. 102(a) Known or Used by Others in This Country Prior to the Applicant’s Invention

In interpreting whether an invention has been known or used in this country, it has been held that the knowledge must consist of a complete and adequate description of the claimed invention and that this knowledge must be available, in some form, to the public. Prior use of an invention in this country by another will be disabling only if the invention in question has actually been reduced to practice and its use has been accessible to the public in some minimal sense. For

Table 1 Summary of the Novelty and Statutory Bar Requirements of 35 U.S.C. 102*Novelty Requirements*

One may not patent an invention if, prior to their date of invention, the invention was any of the following:

1. Known or used by others in this country
2. Patented or described in a printed publication in this or a foreign country
3. Described in a patent granted on an application for patent by another filed in the United States (or in certain international applications)
4. Made in this country by another who had not abandoned, suppressed, or concealed it

Statutory Bar Requirements

One may not patent an invention if one has previously abandoned the invention. One may not patent an invention if, more than one year prior to the time one's patent application is filed, the invention was any of the following:

1. Patented or described in a printed publication in this or a foreign country
2. In public use or on sale in this country
3. Made the subject of an inventor's certificate in a foreign country
4. Made the subject of a foreign patent application, which results in the issuance of a foreign patent before an application is filed in this country

a prior use to be disabling under Section 102(a), the use must have been of a complete and operable product or process that has been reduced to practice.

35 U.S.C. 102(a) Described in a Printed Publication in This or a Foreign Country Prior to the Applicant's Invention

For a printed publication to constitute a full anticipation of a claimed invention, the printed publication must adequately describe the claimed invention. The description must be such that it enables a person of ordinary skill in the art to which the invention pertains to understand and practice the invention. The question of whether a publication has taken place is construed quite liberally by the courts to include almost any act that might legitimately constitute publication. The presence of a single thesis in a college library has been held to constitute publication. Similar liberality has been applied in construing the meaning of the term *printed*.

35 U.S.C. 102(a) Patented in This or a Foreign Country

An invention is not deemed to be novel if it was patented in this country or any foreign country prior to the applicant's date of invention. For a patent to constitute full anticipation and thereby render an invention unpatentable for lack of novelty, the patent must provide an adequate, operable description of the invention. The standard to be applied under Section 102(a) is whether the patent "describes" a claimed invention. A pending patent application is treated as constituting a "patent" for purposes of applying Section 102(a) as of the date of its issuance.

35 U.S.C. 102(e) Described in a Patent Filed in This Country Prior to the Applicant's Invention

Section 102(e) prescribes that if another inventor has applied to protect an invention before you invent the same invention, you cannot patent the invention. The effective date of a U.S. patent, for purposes of Section 102(e) determination, is the filing date of its application, rather than the date of patent issuance. Patent applications that have been published by the USPTO and certain international applications filed under the Patent Cooperation Treaty also are viewed as prior art when applying 35 U.S.C. 102(e).

35 U.S.C. 102(g) Abandoned, Suppressed, or Concealed

For the prior invention of another person to stand as an obstacle to the novelty of one's invention under Section 102(g), the invention made by another must not have been abandoned, suppressed, or concealed. Abandonment, suppression, or concealment may be found when an inventor has been inactive for a significant period of time in pursuing reduction to practice of an invention. This is particularly true when the inventor's becoming active again has been spurred by knowledge of entry into the field of a second inventor.

2.5 Requirement of Utility

To comply with the utility requirement of U.S. patent law, an invention must be capable of achieving some minimal useful purpose that is not illegal, immoral, or contrary to public policy. The invention must be operable and capable of being used for some beneficial purpose. The invention does not need to be a commercially successful product to satisfy the requirement of utility. While the requirement of utility is ordinarily a fairly easy one to meet, problems do occasionally arise with chemical compounds and processes, particularly with pharmaceuticals. An invention incapable of being used to carry out the object of the invention may be held to fail the utility requirement.

2.6 Requirement of Nonobviousness

The purpose of the novelty requirement of 35 U.S.C. 102 and that of the nonobviousness requirement of 35 U.S.C. 103 are the same—to limit the issuance of patents to those innovations that do, in fact, advance the state of the useful arts. While the requirements of novelty and nonobviousness may seem very much alike, the requirement of nonobviousness is a more sweeping one. This requirement maintains that if it would have been obvious (at the time of invention was made) to anyone ordinarily skilled in the art to produce the invention in the manner disclosed, then the invention does not rise to the dignity of a patentable invention and is therefore not entitled to patent protection.

The question of nonobviousness must be wrestled with by patent applicants in the event the USPTO rejects some or all of their claims based on an assertion that the claimed invention is obvious in view of the teaching of one or a combination of two or more prior art references. When a combination of references is relied on in rejecting a claim, the argument the USPTO is making is that it is obvious to combine the teachings of these references to produce the claimed invention. When such a rejection has been made, the burden is on the applicant to establish to the satisfaction of the USPTO that the proposed combination of references would not have been obvious to one skilled in the art at the time the invention was made and/or that, even if the proposed combination of references is appropriate, it still does not teach or suggest the claimed invention.

In an effort to ascertain whether a new development is nonobvious, the particular facts and circumstances surrounding the development must be considered and weighed as a whole. While the manner in which an invention was made must not be considered to negate the patentability of an invention, care must be taken to ensure that the question of nonobviousness is judged as of the time the invention was made and in the light of the then existing knowledge and state of the art.

The statutory language prescribing the nonobviousness requirement is found in Title 35, Section 103, which states:

A patent may not be obtained ... if the differences between the subject matter sought to be patented and the prior art such that the subject matter as a whole would have been obvious at the time the invention was made to a person of ordinary skill in the art to which said subject matter pertains.

In the landmark decision of *Graham v. John Deere*, 383 U.S. 1, 148 U.S.P.Q. 459 (1966), the U.S. Supreme Court held that several basic factual inquiries should be made in determining nonobviousness. These inquiries prescribe a four-step procedure or approach for judging nonobviousness. First, the scope and content of the prior art in the relevant field or fields must be ascertained. Second, the level of ordinary skill in the pertinent art is determined. Third, the differences between the prior art and the claims at issue are examined. Fourth and finally, a determination is made as to whether these differences would have been obvious to one of ordinary skill in the applicable art at the time the invention was made.

2.7 Statutory Bar Requirements

Despite the fact that an invention may be new, useful, and nonobvious and that it may satisfy the other requirements of the patent law, an inventor can still lose the right to pursue patent protection on the invention unless he or she complies with certain requirements of the law called *statutory bars*. The statutory bar requirements ensure that inventors will act with diligence in pursuing patent protection.

While 35 U.S.C. 102 includes both the novelty and the statutory bar requirements of the law (see the summary presented in Table 1), it intertwines these requirements in a complex way that is easily misinterpreted. The novelty requirements are basic to a determination of patentability in the same sense as are the requirements of statutory subject matter, originality, and nonobviousness. The statutory bar requirements are not basic to a determination of patentability but rather operate to decline patent protection to an invention that may have been patentable at one time.

Section 102(b) bars the issuance of a patent if an invention was “in public use or on sale” in the United States more than one year prior to the date of the application for a patent. Section 102(c) bars the issuance of a patent if a patent applicant has previously abandoned the invention. Section 102(d) bars the issuance of a patent if the applicant has caused the invention to be first patented in a foreign country and has failed to file an application in the United States within one year after filing for a patent in a foreign country.

Once an invention has been made, the inventor is under no specific duty to file a patent application within any certain period of time. However, should one of the “triggering” events described in Section 102 occur, regardless of whether this occurrence may have been the result of action taken by the inventor or by actions of others, the inventor must apply for a patent within the prescribed period of time or be barred from obtaining a patent.

Some of the events that trigger statutory bar provisions are the patenting of an invention in this or a foreign country; the describing in a printed publication of the invention in this or a foreign country; the public use of the invention in this country; or putting the invention on sale in this country. Some public uses and putting an invention on sale in this country will not trigger statutory bars if these activities were incidental to experimentation; however, the doctrine of experimental use is a difficult one to apply because of the conflicting decisions issued on this subject.

Certainly, the safest approach to take is to file for patent protection well within one year of any event leading to the possibility of any statutory bar coming into play. If foreign patent protections are to be sought, the safest approach is to file an application in this country before any public disclosure is made of the invention anywhere in the world.

3 PREPARING TO APPLY FOR A PATENT

Conducting a patentability search prior to the preparation of a patent application can be extremely beneficial even when an inventor is convinced that no one has introduced a similar invention into the marketplace.

3.1 Patentability Search

A properly performed patentability study will guide not only the determination of the scope of patent protection to be sought but also the claim-drafting approaches that will be utilized in preparing the patent application. In almost every instance, a patent attorney who has at hand the results of a carefully conducted patentability study can do a better job of drafting a patent application, thereby helping to ensure that it will be prosecuted smoothly, at minimal expense, through the rigors of examination in the USPTO.

Occasionally, a patentability search will indicate that an invention is totally unpatentable. When this is the case, the search will have saved the inventor the cost of preparing and filing a patent application. At times a patentability search turns up one or more newly issued patents that pose infringement concerns. A patentability search is not, however, as extensive a search as might be conducted to locate possible infringement concerns when a great deal of money is being invested in a new product.

Some reasonable limitation is ordinarily imposed on the scope of a patentability search to keep search costs within a relatively small budget. The usual patentability search covers only U.S. patents and does not extend to foreign patents or to publications. Only patents found in the most pertinent USPTO subclasses are reviewed. Even though patentability studies are not of exhaustive scope, a carefully conducted patentability search ordinarily can be relied on to give a decent indication of whether an invention is worthy of pursuing patent coverage to protect, and the information provided by such a search should help one's patent attorney prepare a better patent application than would have been drafted absent knowledge of the search results.

3.2 Putting Invention in Proper Perspective

It is vitally important that a client take whatever time is needed to make certain that his or her patent attorney fully understands the character of an invention before the attorney undertakes the preparation of a patent application. The patent attorney should be given an opportunity to talk with those involved in the development effort from which an invention has emerged. He or she should be told what features these people believe are important to protect. Moreover, the basic history of the art to which the invention relates should be described, together with a discussion of the efforts made by others to address the problems solved by the present invention.

The client should also convey to his or her patent attorney how the present invention fits into the client's overall scheme of development activities. Much can be done in drafting a patent application to lay the groundwork for protection of future developments. Additionally, one's patent attorney needs to know how product liability concerns may arise with regard to the present invention so that statements he or she makes in the patent application will not be used to the client's detriment in product liability litigation. Personal injury lawyers have been known to scrutinize the representations made in a manufacturer's patents to find language that will assist in obtaining recoveries for persons injured by patented as well as unpatented inventions of the manufacturer.

Before preparation of an application is begun, careful consideration should be given to the scope and type of claims that will be included. In many instances, it is possible to pursue both process and product claims. Also, in many instances, it is possible to present claims approaching the invention from varying viewpoints so different combinations of features can be covered. Frequently, it is possible to couch at least two of the broadest claims in different language so efforts of competitors to design around the claim language will be frustrated.

Careful consideration must be given to approaches competitors may take in efforts to design around the claimed invention. The full range of invention equivalents also needs to be taken into account so that claims of appropriate scope will be presented in the patent application.

3.3 Preparing Application

A well-drafted patent application should be a readable and understandable document. If it is not, insist that your patent attorney rework it. A patent application that accurately describes an invention without setting forth the requisite information in a clear and convincing format may be legally sufficient, but it does not represent the quality of work a client has the right to expect.

The claims of a patent application (i.e., the numbered paragraphs that are found at the end of a patent application) are the most important elements of a patent application. All other sections of a patent application must be prepared with care to ensure that these other sections cannot be interpreted as limiting the coverage that is provided by the claims of the application.

A well-drafted patent application should include an introductory section that explains the background of the invention and the character of the problems that are addressed by the invention. It should discuss the closest prior art known to the applicant and should indicate how the invention patentably differs from prior art proposals.

The application should present a brief description of such drawings as form a part of the application, followed by a detailed description of the best mode known to the inventor for carrying out the invention. In the detailed description, one or more embodiments of the invention are described in sufficient detail to enable a person having ordinary skill in the art to which the invention pertains to practice the invention. While some engineering details, such as dimensions, materials of construction, circuit component values, and the like, may be omitted, all details critical to the practice of the invention must be included. If there is any question about the essential character of a detail, prudent practice would dictate its inclusion.

The claims are the most difficult part of the application to prepare. While the claims tend to be the most confusing part of the application, the applicant should spend enough time wrestling with the claims and/or discussing them with the patent attorney to make certain that the content of the claims is fully understood. In drafting the claims of an application, legal gibberish should be avoided, such as endless uses of the word *said*. Elements unessential to the practice of the invention should be omitted from the broadest claims, and essential elements should be described in the broadest possible terms.

The patent application will usually include one or more sheets of drawings and will be accompanied by a suitable declaration or oath to be signed by the inventor or inventors. The drawings of a patent application should illustrate each feature essential to the practice of the invention and show every feature to which reference is made in the claims. The drawings must comply in size and format with a lengthy set of technical rules promulgated and frequently updated by the USPTO. The preparation of the drawings is ordinarily best left to an experienced patent draftsman.

A well-prepared application will help to pave the way for smooth handling of the patent application during its prosecution. If a patent application properly tells the story of the invention, it should constitute something of a teaching document that will stand on its own and be capable of educating a court regarding the character of the art to which the invention pertains as well as the import of this invention to that art. Since patent suits are tried before judges who rarely have technical backgrounds, it is important that a patent application make an effort to present the basic features of the invention in terms understandable by those having no technical training. It is unusual for an invention to be so impossibly complex that its basic thrust defies description in fairly simple terms. A patent application is suspect if it wholly fails to set forth, at some point, the pith of the invention in terms a grade school student can grasp.

3.4 Enablement, Best Mode, Description, and Distinctness Requirements

Once a patent application has been prepared and is in the hands of the inventor for review, it is important that the inventor keep in mind the enablement, best mode, description, and distinctness requirements of the patent law. The enablement requirement calls for the patent application

to present sufficient information to enable a person skilled in the relevant art to make and use the invention. The disclosure presented in the application must be such that it does not require one skilled in the art to experiment to any appreciable degree to practice the invention.

The best mode requirement mandates that an inventor disclose, at the time he or she files a patent application, the best mode he or she then knows about for practicing the invention.

The description requirement also relates to the descriptive part of a patent application and the support it must provide for any claims that may need to be added after the application has been filed. Even though a patent application may adequately teach how to make and use the subject matter of the claimed invention, a problem can arise during the prosecution of a patent application where one determines it is desirable to add claims that differ in language from those filed originally. If the claim language one wants to add does not find adequate support in the originally filed application, the benefit of the original filing date will be lost with regard to the subject matter of the claims to be added—a problem referred to as *late claiming*, about which much has been written in court decisions of the past half century. Therefore, in reviewing a patent application prior to its being executed, an inventor should keep in mind that the description that forms a part of the application should include support for any language he or she may later want to incorporate in the claims of the application.

The distinctness requirement applies to the content of the claims. In reviewing the claims of a patent application, an inventor should endeavor to make certain the claims particularly point out and distinctly claim the subject matter that he or she regards as his or her invention. The claims must be *definite* in the sense that their language must clearly set forth the area over which an applicant seeks exclusive rights. The language used in the claims must find antecedent support in the descriptive portion of the application. The claims must not include within their scope of coverage any prior art known to the inventor and yet should present the invention in the broadest possible terms that patentably distinguish the invention over the prior art.

3.5 Product-by-Process Claims

In some instances, it is possible to claim a product by describing the process or method of its manufacture. Some products are unique because of the way they are produced; hence securing proper protection may necessitate that the claims of the patent application recite critical process steps that cause the resulting product to be unique. Patentability cannot be denied due to the way a product is made but may be enhanced if the process lends novelty to the resulting product.

3.6 Claim Format

Patent applicants have some freedom in selecting the terminology that they use to define and claim their inventions, for it has long been held that “an applicant is his own lexicographer.” However, the meanings that the applicants assign to the terminology they use must not be repugnant to the well-known usages of such terminology. When an applicant does not define the terms he or she uses, such terms must be given their “plain meaning,” namely the meanings given to such terms by those of ordinary skill in the relevant art.

Each claim is a complete sentence. In many instances, the first part of the sentence of each claim appears at the beginning of the claims section and reads, “What is claimed is:” Each claim typically includes three parts: preamble, transition, and body. The preamble introduces the claim by summarizing the field of the invention, its relation to the prior art, and its intended use or the like. The transition is a word or phrase connecting the preamble to the body. The terms *comprises* and *comprising* often perform this function. The body is the listing of elements and limitations that define the scope of what is being claimed.

Claims are either *independent* or *dependent*. An independent claim stands on its own and makes no reference to any other claim. A dependent claim refers to another claim that may be

independent or dependent and adds to the subject matter of the reference claim. If a dependent claim depends from (makes reference to) more than one other claim, it is called a *multiple dependent* claim.

One type of claim format that can be used gained notoriety in a 1917 decision of the Commissioner of Patents, *Ex parte Jepson*, 1917 C.D. 62. In a claim of the Jepson format, the preamble recites all the elements deemed to be old, the body of the claim includes only such new elements as constitute improvements, and the transition separates the old from the new. The USPTO favors the use of Jepson-type claims since this type of claim is thought to assist in segregating what is old in the art from what the applicant claims as his or her invention.

In 1966, the USPTO sought to encourage the use of Jepson-type claims by prescribing the following rule 75(e):

Where the nature of the case admits, as in the case of an improvement, any independent claim should contain in the following order, (1) a preamble comprising a general description of the elements or steps of the claimed combination which are conventional or known, (2) a phrase such as “wherein the improvement comprises,” and (3) those elements, steps and/or relationships which constitute that portion of the claimed combination which the applicant considers as the new or improved portion.

Thankfully, the use of the term *should* in Rule 75(e) makes use of Jepson-type claims permissive rather than mandatory. Many instances occur when it is desirable to include several distinctly old elements in the body of the claim. The preamble in a Jepson-type claim has been held to constitute a limitation for purposes of determining patentability and infringement, while the preambles of claims presented in other types of format may not constitute limitations.

The royalties one can collect from a licensee often depend on how the claims of the licensed patent are worded. A 2% royalty on a threaded fastener used in a locomotive amounts to much less than a 2% royalty on a locomotive that is held together by the novel fastener; therefore, it may be wise for the patent to include claims directed to the locomotive instead of limiting all of the claims to features of the relatively inexpensive fastener. A proper understanding of the consequences of presenting claims in various formats, and of the benefits thereby obtained, should be taken into account by one’s patent attorney.

3.7 Executing Application

Once an inventor is satisfied with the content of a proposed patent application, he or she should read carefully the oath or declaration accompanying the application. The required content of this formal document recently has been simplified. In it, the inventor states that he or she:

1. Has reviewed and understands the content of the application, including the claims, as amended by any amendment specifically referred to in the oath or declaration.
2. Believes the named inventor or inventors to be the original and first inventor or inventors of the subject matter which is claimed and for which a patent is sought.
3. Acknowledges the duty to disclose to the USPTO during examination of the application all information known to the person to be material to patentability.

If the application is being filed as a division, continuation, or continuation-in-part of one or more co-pending patent applications, the parent case or cases must be adequately identified in the oath or declaration. Additionally, if a claim to the benefit of a foreign-filed application is being made, the foreign-filed application must be adequately identified in the oath or declaration.

Absolutely no changes should be made in any part of a patent application once it has been executed. If some change, no matter how ridiculously minor, is found to be required after an application has been signed, the executed oath or declaration must be destroyed and a new one

signed after the application has been corrected. If an application is executed without having been reviewed by the applicant or is altered after having been executed, it may be stricken from the files of the USPTO.

3.8 U.S. Patent and Trademark Office Fees

The USPTO charges a set of fees to file an application, a fee to issue a patent, fees to maintain a patent if it is to be kept alive for its full available term, and a host of other fees for such things as obtaining an extension of time to respond to an office action. The schedule of fees charged by the USPTO is updated periodically, usually resulting in fee increases. Such fee increases often take effect on or about October 1, when the government's new fiscal year begins. This has been known to result in increased numbers of September filings of applications during years when sizable fee increases have taken effect.

In addition to a basic filing fee of \$300, \$200 is charged for each independent claim in excess of a total of three, \$50 is charged for each claim of any kind in excess of a total of 20, and \$360 is charged for any application that includes one or more multiple dependent claims. Also, there is a search fee of \$500 as well as an examination fee of \$200 that must be paid either when an application is filed or shortly thereafter. However, if the applicant is entitled to claim the benefit of so-called *small-entity status*, all of these fees are halved, as are most other fees that are associated with the prosecution and issuance of a patent application.

Provisional applications require a \$200 filing fee that may be halved for small entities. Applications for design patents require a filing fee of \$200, a search fee of \$100, and an examination fee of \$130, unless small-entity status is claimed, whereupon these fees also may be halved. Plant patent applications require a \$300 filing fee, a \$300 search fee, and a \$160 examination fee that are halved for small entities.

New rules now permit the USPTO to assign a filing date before the filing fee and oath or declaration have been received. While the filing fee and an oath or declaration are still needed to complete an application, a filing date will now be assigned as of the date of receipt of the descriptive portion of an application (known as the specification) accompanied by at least one claim, any required drawings, and a statement of the names and citizenships of the inventors.

The issue fee charged by the USPTO for issuing a utility patent on an allowed application stands at \$1400. Establishing a right to the benefits of small-entity status permits reduction of this fee to \$700. The issue fee for a design application is \$800, which also may be halved with the establishment of small-entity status. A plant patent requires an issue fee of \$1100, which may be halved for small entities. There is no issue fee associated with a provisional application since a provisional application does not issue as a patent unless it is supplemented within one year of its filing date by the filing of a complete utility application.

Maintenance fees must be paid to keep an issued utility patent in force during its term. No maintenance fees are charged on design or plant patents or on utility patents that have issued from applications filed before December 12, 1980. As of this writing, maintenance fees of \$900, \$2300, and \$3800 are due no later than $3\frac{1}{2}$, $7\frac{1}{2}$, and $11\frac{1}{2}$ years, respectively, from a utility patent's issue date. Qualification for the benefits of small-entity status allows these fees to be reduced to \$450, \$1150, and \$1900, respectively. Failure to timely pay any maintenance fee or to late pay it during a six-month grace period following its due date accompanied by a late payment surcharge of \$130 (\$65 for small entities) will cause a patent to lapse permanently.

3.9 Small-Entity Status

Qualification for the benefit of small-entity status requires only the filing of a claim to the right-to-pay fees at a small-entity rate. All entities having rights with respect to an application or patent must each be able to qualify for small-entity status; otherwise, small-entity status

cannot be achieved. Once a claim to the benefit of small-entity status has been presented to the USPTO, there exists a continuing duty to advise the USPTO before or at the time of paying the next fee if qualification for small-entity status has been lost.

Those who qualify for small-entity status include:

1. A sole inventor who has not transferred his or her rights and is under no obligation to transfer his or her rights to an entity that fails to qualify
2. Joint inventors where no one among them has transferred his or her rights and is under no obligation to transfer his or her rights to an entity that fails to qualify
3. A nonprofit organization such as an institution of higher education or an IRS-qualified and exempted nonprofit organization
4. A small business that has no more than 500 employees after taking into account the average number of employees (including full time, part time, and temporary) during the fiscal year of the business entity in question and of its affiliates, with the term *affiliate* being defined by a broad-reaching “control” test

Attempting to establish small-entity status fraudulently or claiming the right to such status improperly or through gross negligence is considered a fraud on the USPTO. An application could be disallowed for such an act.

Failure to claim small-entity status on a timely basis may forfeit the right to small-entity status benefits with respect to a fee being paid. However, if a claim to small-entity status is presented to the USPTO within two months after a fee was paid, a refund of the excess amount paid may be obtained.

3.10 Express Mail Filing

During 1983, a procedure was adopted by the USPTO that permits certain papers and fees to be filed in the USPTO by using the “Express Mail Post Office to Addressee” service of the U.S. Postal Service. When this is done, the filing date of the paper or fee will be the mailing date affixed to the “Express Mail” mailing label by USPS personnel.

To qualify for the filed-when-mailed advantage, each paper must bear the number of the “Express Mail” mailing label, must be addressed to the Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, and must comply with the other requirements that change from time to time.

4 PROSECUTING PENDING PATENT APPLICATION

Once an executed patent application has been received by the USPTO, the patent application is said to be pending, and a serial number and filing date are assigned to it by the USPTO. The prosecution period of a patent application is the time during which an application is pending. It begins when a patent application is filed in the USPTO and continues until either a patent is granted or the application is abandoned. Activities that take place during this time are referred to as *prosecution*.

4.1 Patent Pending

Shortly after an application is filed, a filing receipt is sent to the applicant or the applicant’s attorney. The receipt provides evidence of filing and sets out important data summarizing what the USPTO received (i.e., what was filed in the USPTO). It also lists the assigned serial number, filing date, application title, and inventor identification.

Once an application for a patent has been received by the USPTO, the applicant may mark products embodying the invention and literature or drawings relating to the invention with an indication of “Patent Pending” or “Patent Applied For.” These expressions mean a patent application has been filed and has neither been abandoned nor issued as a patent. The terms do not mean that the USPTO has taken up examination of the merits of an application, much less approved the application for issuance as a patent.

Marking products with the designation “Patent Pending” often has the effect of discouraging competitors from copying an invention, whereby the term of the patent that eventually issues may be thought of as effectively extended to include the period during which the application is pending. In many instances, competitors will not risk a substantial investment in preparation for the manufacture and merchandising of a product when a competing product is marked with the designation “Patent Pending,” for they know their efforts may be legally interrupted as soon as the patent issues.

4.2 Publication of Pending Applications

In an effort to harmonize U.S. patent practice with that of other countries, the USPTO now is “publishing” pending utility patent applications 18 months after each application is filed. If the application claims the benefit of the filing date of an earlier filed application, then the 18-month publication date is measured from the first filing date that provides benefit. Early publication also can be requested by a patent applicant, but a number of formalities must be met and a fee must be paid if applicant-requested early publication is to take place.

A patent applicant can request that his or her application not be published, but the USPTO will honor this request only if the applicant presents it to the USPTO at the time the application is filed, and if the request is accompanied by a certification stating that the invention is not and will not be the subject of foreign or international applications (i.e., no foreign filing of corresponding cases will be made). If, after filing a nonpublication request, the applicant changes his or her mind and files abroad, then the applicant must notify the USPTO within 45 days so that publication can proceed on schedule in the USPTO.

To compensate patent applicants for a loss of secrecy when the content of pending applications is published, so-called provisional rights are provided by law that can be asserted against anyone practicing inventions claimed in a pending and published application for infringements that take place between the date of publication of the application and the date of issuance of the patent, so long as at least one infringed claim of the patent is “substantially identical” to a claim found in the published application.

Another factor to take into account in deciding whether to request nonpublication of a ready-to-file patent application is that, once the application is published, the file of the application, which, until publication, has been held in secrecy by the USPTO, becomes open to public inspection upon payment of a fee by someone who desires to inspect the file. When the USPTO “publishes” a pending utility patent application, they do it entirely electronically: No paper copies are published or distributed. To view a published pending application, one must visit the USPTO website www.uspto.gov, which provides a link to a search utility where information can be entered to search for published applications using key words, inventor names, or other pertinent data. Thus, the files of published pending applications are open to the public to virtually the same extent as are the files of issued patents, and members of the public can submit prior art and other information to the USPTO to be considered as the prosecution of the published application is brought to a close.

Applicants who want the USPTO to keep secret their pending utility application until the day it may issue must make that choice before they file the application and must elect nonpublication at the time when they file the application in the USPTO. The USPTO will not honor a request made to maintain an application in confidence if the request is made at any time after the application is filed.

4.3 Duty of Candor

It is extremely important for patent applicants to deal candidly with the USPTO. In accordance with USPTO guidelines, a patent applicant is urged to submit an information disclosure statement either concurrently with the filing of an application or within three months of its filing. An information disclosure statement should include a listing of patents, publications, or other information that is believed to be “material” and a concise explanation of the relevance of each listed item and should be accompanied by copies of each listed reference that is not a U.S. patent. Items are deemed to be “material” where there is a “substantial likelihood that a reasonable examiner would consider it important in deciding whether to allow the application to issue as a patent.”

To ensure that the USPTO will give due consideration to what is cited in an information disclosure statement, the information disclosure statement must be filed (1) within three months of the filing date of a normal U.S. application, or (2) within three months of entry of a U.S.-filed international application into its national stage, or (3) before the mailing date of a first office communication treating the merits of the claimed invention (known as an “office action”), whatever occurs last. Consideration thereafter can be had only if other requirements are met, which typically include the certification of certain information, the filing of a petition, and/or the payment of a \$180 fee. Information disclosure statements filed before the grant of a patent that do not comply with the requirements of the USPTO are not considered by the USPTO but will be placed in the official file of the patent.

The courts have held that those who participate in proceedings before the USPTO have the “highest duty of candor and good faith.” While the courts differ in their holding of the consequences of misconduct, fraud on the USPTO has been found to be a proper basis for taking a wide variety of punitive actions, such as striking applications from the records of the USPTO, canceling issued patents, denying enforcement of patents in infringement actions, awarding attorney’s fees to defendants in infringement actions, and imposing criminal sanctions on those who were involved in fraudulently procuring patents. Inequitable conduct other than outright fraud has been recognized as a defense against enforcement of a patent, as a basis for awarding attorney’s fees in an infringement action, and as a basis of antitrust liability.

In short, the duty of candor one has in dealings with the USPTO should be taken very seriously. Prudent practice would urge that if there is any question concerning whether a reference or other facts are “material,” a citation should be made promptly to the USPTO so that the examiner can decide the issue.

4.4 Initial Review of Application

Promptly after an application is filed, it is examined to make certain it is complete and satisfies formal requirements sufficiently to permit its being assigned a filing date and serial number. The classification of the subject matter of the claimed invention is determined and the application is assigned to the appropriate examining group. In the group, the application is assigned to a particular examiner. Examiners take up consideration of the applications assigned to them in the order of their filing.

Although more than 2000 examiners staff the USPTO, a backlog of several months of cases awaits action. This results in a delay of several months between the time an application is filed and when it receives its first thorough examination on the merits. Most newly filed patent applications receive a first office action treating the merits of their invention within about 14 months after the application was filed.

Once an examiner reaches an application and begins the initial review, he or she checks the application still further for compliance with formal requirements and conducts a search of the prior art to determine the novelty and nonobviousness of the claimed invention. The

examiner prepares an office action, in which he or she notifies the applicant of any objections to the application or requirements regarding election of certain claims for present prosecution, and/or any rejections he or she believes should be made of the claims. In some instances, the examiner will find it necessary to object to the form of the application. One hopes that these formal objections are not debilitating and can be corrected by relatively minor amendments made in response to the office action.

In treating the merits of the claims, especially in the first office action, it is not uncommon for an examiner to reject the majority, if not all, of the claims. Some examiners feel strongly that they have a duty to cite the closest art they are able to find and to present rejections based on this art to encourage or force the inventor to put on record in the file of the application such arguments as are needed to illustrate to the public exactly how the claimed invention distinguishes patentably over the cited art.

In the event the examiner deems all the claims in an application to be patentable, he or she notifies the applicant, typically by issuing a notice of allowance.

4.5 Response to Office Action

Unless an applicant can show that his or her delay in responding to an office action was “unavoidable,” the applicant must respond to office actions within six months of their mailing dates (or sooner if the office action sets earlier deadlines). Failure to respond on a timely basis causes the application to become “abandoned.” Most office actions set a two- or three-month deadline for responding that is measured from their mailing dates. If an extension of time is needed to respond, extensions can be obtained upon petition and payment of the requisite fee, but extensions cannot move the ultimate response deadline beyond six months from the date of mailing of an office action. The fees for one-, two-, and three-month extensions currently are \$110, \$430, and \$980, respectively. Qualified small entities pay fees that are one-half of these amounts.

In the event the first office action issued by the examiner is adverse in any respect and/or leaves one or more issues unresolved, the applicant may reply in a variety of ways that constitute a bona fide attempt to advance the prosecution of the application. The applicant is entitled to at least one reconsideration by the USPTO following the issuance of the first office action; however, as a minimum, a response must present at least some argument or other basis for requesting reconsideration.

The rejection of one or more claims can be responded to by the presentation of arguments to persuade the examiner that the examiner’s position is inappropriate or incorrect or fails to take into account important information; by amending the rejected claims to overcome or to sidestep the issues raised by the examiner; by submitting affidavits or declarations presenting additional information that the examiner should consider; by canceling the rejected claim or claims; and/or by adding new claims and arguing for their allowance. In responding to an office action, each objection and rejection made by the examiner must be treated. If the inventor agrees that certain of his or her claims should not be allowed in view of art cited by the examiner, these claims may be canceled or amended to better distinguish the claimed invention over the cited art.

Since the file of a patent application will become open to public inspection on issuance of a patent, and because an issued patent must be interpreted in view of the content of its file, the character of any arguments presented to the USPTO in support of a claimed invention are critical. Care must be taken in the drafting of arguments to ensure that no misrepresentations are made and that the arguments will not result in an unfavorable interpretation of allowed claims being made during the years when the resulting patent is in force. Great care also must be taken in drafting claim amendments, for any “narrowing” amendment will probably cause the amended claims to be interpreted very narrowly.

Years ago, it was not unusual for half a dozen or more office actions to be generated during the course of pendency of a patent application. During recent years, however, the USPTO has placed emphasis on “compacting” the prosecution of patent applications and insists that responses to office actions make a genuine, full-fledged effort to advance the prosecution of the application. Today, it is not unusual for the prosecution of a patent application to be concluded on the issuance of the second or third office action.

If the USPTO has objected to the drawings, corrections must be made by providing substitute drawings that include the required corrections. The earlier that substitute drawings can be submitted by the applicant, the better.

4.6 Reconsideration in View of Filing of Response

Once the applicant has responded to an office action, the examiner reexamines the case and issues a new office action apprising the applicant of his or her findings. If the examiner agrees to allow all of the claims that remain active in the application, prosecution on the merits is closed and the applicant may not present further amendments or add other claims as a matter of right. If the office action is adverse with regard to the merits of the claims, the prosecution of the case continues until such time as the examiner issues an office action that presents a final rejection.

A rejection is made “final” once a clear and unresolved issue has developed between the examiner and the applicant. After a final rejection has issued, the character of the responses that may be made by the applicant is limited. The applicant may appeal the final rejection to an intraagency Board of Patent Appeals and Interferences, cancel the rejected claims, comply with all the requirements for allowance if any have been laid down by the examiner, or file a request for continued examination (RCE) whereby the examination procedure is begun again.

If an initial appeal taken to the Board of Patent Appeals and Interferences should result in an unfavorable decision, a further appeal may be taken to either the U.S. District Court for the District of Columbia or to the U.S. Court of Appeals for the Federal Circuit. In some instances, further appeals may be pursued to a higher court.

In the majority of instances during the period of prosecution, the application eventually reaches a form acceptable to the examiner handling the application, and the examiner will issue a notice of allowance. If it is impossible to reach accord with the examiner handling the application, the inventor can make use of the procedures for appeal or can continue the prosecution of the application by filing an RCE.

If the record of examination of an application does not otherwise reveal the reasons for allowance, an examiner may put a comment in the file explaining his or her reasons for allowing the claims that have been allowed. If a statement of reasons for allowance is provided by an examiner, it should be reviewed with care and commented upon, in writing, if the stated reason is incorrect or needs clarification.

4.7 Interviewing Examiner

If, during the prosecution of a patent application, it appears that substantial differences of opinion or possible misunderstandings are being encountered in dealing with the examiner to whom the application has been assigned, it often is helpful for the attorney to conduct a personal interview with the examiner. While the applicant has a right to attend such a meeting, this right is best exercised sparingly and usually requires that the applicant spend time with the attorney to become better prepared to advance rather than to detract from the presentation the attorney plans to make.

Considering the relatively sterile and terse nature of many office actions, it may prove difficult to determine accurately what the examiner’s opinion may be regarding how the application should be further prosecuted. If it has become clear that an examiner and an attorney are

not communicating in the full sense of the word, a personal interview often will prove to yield valuable guidance for bringing the prosecution of the application to a successful conclusion. In other instances, an interview will be beneficial in more correctly ascertaining the true character of any difference of opinion between the applicant and the examiner, thereby enabling the exact nature of remaining obstacles to be addressed thoroughly in the applicant's next response.

4.8 Restriction and Election Requirements

If a patent examiner determines that an application contains claims to more than one independent and distinct invention, the examiner may impose what is called a *restriction requirement*. In the event the examiner finds that the application claims alternative modes or forms of an invention, he or she may require the applicant to elect one of these species for present prosecution. This is called a *species election requirement*.

Once a restriction or election requirement has been imposed, the applicant must elect one of the designated inventions or species for present prosecution in the original application. The applicant may file divisional applications on the nonelected inventions or species any time during the pendency of the original application, which often results in a plurality of related patents issuing on different aspects of what the inventor regards as a single invention.

When responding to an office action that includes a restriction and/or election requirement, arguments may be presented in an effort to traverse the requirement and request its reconsideration. After traversing, the examiner is obliged to reconsider the requirement, but he or she may repeat it and make it final. Sometimes, the examiner can be persuaded to modify or withdraw a restriction and/or election requirement, thereby permitting a larger number of claims to be considered during the prosecution of the pending application.

4.9 Double-Patenting Rejections

Occasionally, one may receive a rejection based on the doctrine of double patenting. This doctrine precludes the issuance of a second patent on the same invention already claimed in a previously issued patent.

One approach to overcoming a double-patenting rejection is to establish a clear line of demarcation between the claimed subject matter of the second application and that of the earlier patent. If the line of demarcation is such that the claimed subject matter of the pending application is nonobvious in view of the invention claimed in the earlier patent, no double-patenting rejection is proper.

If the claimed subject matter amounts to an obvious variation of the claimed invention of the earlier issued patent, the double-patenting rejection often may be overcome by the filing of a terminal disclaimer. With a terminal disclaimer, a portion of the term of any patent issuing on the pending application is disclaimed so that any new patent issuing on the pending application will expire on the same day the already existing patent expires. If, however, the claimed subject matter of the pending application is identical to the claimed subject matter in the earlier issued patent, it is not possible to establish a line of demarcation between the two cases and the pending application is not patentable even if a terminal disclaimer is filed.

4.10 Continuation, Divisional, and Continuation-in-Part Applications

During the pendency of an application, it may be desirable to file either a continuation or a divisional application. A divisional application may be filed when two or more inventions are disclosed in the original application and claims to only one of these are considered during examination of the originally filed case.

It frequently occurs during the pendency of a patent application that a continuing program of research and development being conducted by the inventor results in the conception of improvements in the original invention. Because of a prohibition in the patent law against amending the content of a pending application to include “new matter,” any improvements made in the invention after the time an application is filed cannot be incorporated into a pending application. When improvements are made that are deemed to merit patent protection, a continuation-in-part application is filed. Such an application can be filed only during the pendency of an earlier filed application commonly called the *parent case*. The continuation-in-part case receives the benefit of the filing date of the parent case with regard to such subject matter as is common to the parent case. However, any subject matter that is not in common with the parent case is entitled only to the benefit of the filing date of the continuation-in-part case.

In some instances, when a continuation-in-part application has been filed, the improvements that form the subject matter of the continuation-in-part case are closely associated with the subject matter of the earlier filed application, and the earlier application may be deliberately abandoned in favor of the continuation-in-part case. In other instances, the new matter that is the subject of the continuation-in-part application clearly constitutes an invention in and of itself. In such a situation, it may be desirable to continue the prosecution of the original application to obtain one patent that covers the invention claimed in the original application and a second patent that covers the improvement features.

4.11 Maintaining Chain of Pending Applications

If a continuing development program is underway that produces a series of improvements, it can be highly advantageous to maintain on file in the USPTO a continuing series of pending applications—an unbroken chain of related cases. If an original parent application is initially filed and a series of continuation, division, and/or continuation-in-part applications are filed in a manner that ensures the existence of an uninterrupted chain of pending cases, any patent or patents that may issue on the earlier cases cannot be used as references cited by the USPTO as obstacles in the path of allowance of later applications in the chain. This technique of maintaining a series or chain of pending applications is an especially important technique to use when the danger exists that the closest prior art the USPTO may be able to cite against the products of a continuing research and development effort is the patent protection that issued on early aspects of this effort.

4.12 Patent Issuance

Once a notice of allowance has been mailed by the USPTO, the applicant has an inextensible period of three months to pay the issue fee. The notice of allowance usually is accompanied by a paper that advises whether the normal term of the patent is to be adjusted to compensate for delays caused during application prosecution by either the applicant or the USPTO. The applicant can present arguments for lengthening the term of the patent if he or she believes that the term adjustment should be modified.

Payment of the issue fee is a prerequisite to the issuance of a patent. If payment of the issue fee is unavoidably or unintentionally late, the application becomes abandoned but usually can be revived within a year of the payment due date. Reviving an unintentionally abandoned application requires a much higher fee payment than does revival of an unavoidably abandoned application. A patent will not issue unless this fee is paid.

A few weeks before the patent issues, the USPTO mails a notice of issuance, which advises the applicant of the issue date and patent number.

Upon receipt of a newly issued patent, it should be reviewed with care to check for printing errors. If printing errors of misleading or otherwise significant nature are detected, it is desirable

to petition for a certificate of correction. If errors of a clerical or typographical nature have been made by the applicant or by his or her attorney and if these errors are not the fault of the USPTO, a fee must be paid to obtain the issuance of a certificate of correction. If the errors are the fault of the USPTO, no such fee need be paid.

The issuance of a patent carries with it a presumption of validity. As was stated by Judge Markey in *Roper Corp. v. Litton Systems, Inc.*, 757 F.2d 1266 (Fed. Cir., 1985), "A patent is born valid and remains valid until a challenger proves it was stillborn or had birth defects." If the validity of a patent is put in question, the challenger has the burden of establishing invalidity by evidence that is clear and convincing.

4.13 Safeguarding Original Patent Document

The original patent document merits appropriate safeguarding. It is printed on heavy bond paper, its pages are fastened securely together, and it bears the official seal of the USPTO. The patent owner should preserve this original document in a safe place as evidence of his or her proprietary interest in the invention. If an infringer must be sued, the patent owner may be called on to produce the original patent document in court.

4.14 Reissue

The applicant or owner of a patent may apply for a reissue patent to correct errors made without deceptive intent that cause a patent to be wholly or partly inoperative or invalid because of a defective specification or drawing or because the patentee claimed more or less than he had a right to claim. A reissue application may be filed in the USPTO to correct an obvious error in a specification, add a priority claim to obtain the benefit of the filing date of an earlier filed application, cure claim indefiniteness, correct a misdesignation of inventors, broaden or narrow the scope of claims, and correct other errors resulting from an inadvertent or accidental mistake. If a patentee seeks to enlarge by reissue the scope of the claims of the original patent, he or she must file a broadening reissue application with two years of the date of issuance of the original patent. If the patentee seeks to narrow his claims or otherwise correct some defect in a patent through reissue (other than enlarging the scope of the claims of the patent as discussed in the preceding sentence) he or she may file a narrowing reissue application at any time during the term of the patent.

The filing fee for a reissue application is the same as the filing fee for an original application that has the same content. At least one error or defect must be stated in the oath or declaration that accompanies a reissue application. For an error to be correctable through reissue, it must have arisen without deceptive intention. A patentee may not seek by the route of reissue to change the claims in an issued patent so as to recapture subject matter that was intentionally surrendered to obtain the original patent. Moreover, the reissue application may not contain any new matter and must be directed to what was disclosed in the original patent.

A reissue application is examined in much the same way as original applications. If the grant of a reissue patent is approved, a reissue patent issued by the USPTO will replace the original patent and will expire on the same date that the original patent was set to expire.

4.15 Reexamination

Any person, including the owner of a patent and accused infringers, may file a request for reexamination of the validity of any claim in a patent. The request may be based on prior art or other facts not previously considered by the USPTO that the reissue request brings to the attention of the USPTO. The fee for filing a so-called *ex parte* reexamination request is \$2520, and no small-entity reduction is applicable. The fee for filing an *inter partes* reexamination that

permits the filing entity to participate in the reexamination process is \$8800, and no small-entity reduction is applicable. If the USPTO decides that a reexamination request properly presents a new patentability question, it will issue a reexamination order.

Unlike *ex parte* reexamination, *inter partes* reexamination does not allow a challenger to remain anonymous. However, the two types of reexamination are similar in that each may consider the same type of prior art, and in each type the challenger must explain the significance of the prior art being cited as justifying the narrowing or cancellation of claims found in the original patent. In both types, the USPTO makes the ultimate decisions regarding patentability.

The reexamination of a patent provides the requesting entity with an opportunity to challenge the scope and utility of a patent without having to go to court to do so. Reexaminations are conducted by the USPTO, and claim amendments or cancellations are attended to in much the same way they are handled during prosecution of a normal application.

If an *inter partes* reexamination is conducted, interested third parties may participate on an adversarial basis by submitting arguments and by making proposals for the USPTO to consider. The only restriction on when an *inter partes* reexamination can take place is that a patent must be in existence (reexamination cannot take place until a patent actually issues). Through *inter partes* reexamination, an accused infringer, or a licensee of a patent, or any other interested person can attempt to narrow or invalidate a patent.

Once the reexamination of a patent is completed, a reexamination certificate is published by the USPTO which cancels any original claim of the subject patent determined to be unpatentable, confirms any original claim determined to be patentable, and incorporates into the patent any new claim determined to be patentable.

5 ENFORCING PATENTS AGAINST INFRINGERS

Patent rights cannot be enforced against infringers until a patent has issued. Beginning on the very day of patent issuance, however, efforts can be set in motion to curtail infringements by others and/or seek payments from them for the use of the patented invention.

5.1 Patent Infringement

The law recognizes several types of patent infringement. *Direct infringement* involves the making, using, offering for sale or selling within the United States, or importing into the United States of the entirety of an invention defined by a claim of the patent during the term of the patent. Since the implementation of the Process Patent Amendments Act of 1988, direct infringement also now includes making or using in the United States a product made by a process that is patented in the United States, which applies not only to domestic-origin but also to foreign-origin products. *Inducement of infringement* includes a number of activities by which one may intentionally cause, urge, encourage, or aid another to infringe a patent. *Contributory infringement* occurs when a person aids or abets direct infringement, as is set out in Section 271(c) of Title 35, which states:

Whoever offers to sell, or sells within the United States, a component of a patented machine, manufacture, combination or composition, or a material or apparatus for use in practicing a patented process, constituting a material part of the invention, knowing the same to be especially made or especially adapted for use in an infringement of such patent, and not a staple article or commodity of commerce suitable for substantial non-infringing use, shall be liable as a contributory infringer.

A patent infringement suit can be brought only in a federal district court where either (1) the infringer does business and infringement is committed or (2) the infringer resides. The suit must be brought within a six-year statute of limitations. An infringement action is initiated

by filing a written complaint with the clerk of the district court where the suit is brought. A copy of the complaint is served on the defendant by the court. Either party in an infringement action may demand a jury or the case can be tried by the judge if both parties waive their jury rights. During the suit, the plaintiff has the burden of establishing by competent evidence that the defendant's activities infringe the patent in question. In addition, the plaintiff must defend against any assertion made by the defendant that the plaintiff's patent is invalid.

The first months of an infringement suit usually bring extensive discovery, which involves demands for responses to written questions, document production, and witness depositions. Reports of expert witnesses are exchanged, and the experts are deposed. Either side may move for summary judgment, seeking a court ruling that will avoid the need for trial. A Markman hearing that may involve expert testimony is customary to determine claim interpretation. The trial itself may be divided into separate treatments of infringement and validity. If the plaintiff prevails on liability, the question of damages is then addressed.

Patent litigation is almost always expensive. Both the plaintiff and the defendant in a patent infringement case have to establish complex positions of fact and law. It is not unusual for each of the parties to a patent infringement suit to incur costs of at least a quarter million dollars, especially if the case is tried through appeal. For this reason, patent infringement litigation may not be economical unless there is a market at stake that is considerably larger than the anticipated costs of litigation.

5.2 Defenses to Patent Enforcement

Noninfringement is one of several traditional defenses an accused infringer can raise in his or her behalf. Three other substantive defenses are raised quite commonly in an effort to preclude the enforcement of a patent. The first of these is the defense of patent invalidity. While a patent is presumed to be valid, a court will normally inquire into its validity if the defendant puts the issue in question. Arguments favoring invalidity can be based on a variety of grounds, including the invention's lack of novelty, its obviousness, or insufficient disclosure of the invention in the patent being asserted. In view of the presumption of validity that arises with patent issuance, the relatively stringent standard of "clear and convincing evidence" must be met to establish patent invalidity.

A second, commonly asserted defense is that the patent is unenforceable as a result of fraudulent procurement or inequitable conduct by the patentee before the USPTO. The patent may be rendered unenforceable and the patent owner subject to other liabilities if the defendant can show by clear, unequivocal, and convincing evidence that a breach of the applicant's duty of candor has taken place through an intentional or grossly negligent misrepresentation or withholding of material fact. However, the type of showing required to establish fraud must exceed mere evidence of simple negligence, oversight, or erroneous judgment made in good faith.

Another commonly asserted defense is that of unenforceability due to so-called patent misuse and/or violation of the antitrust laws. If a patent owner has exploited the patent in an improper way by violating the antitrust laws or by effectively extending the patent beyond its lawful scope, the courts will refrain from assisting the patent owner in remedying infringement until the misuse is purged and its consequences dissipated.

5.3 Outcome of Suit

If a patent owner successfully prevails in an infringement suit, the patent owner will be entitled to recover such damages as the patent owner can prove to have suffered, but not less than a reasonable royalty. In some situations, the patent owner may succeed in recovering the total amount of profits that would have resulted from the additional increment of business the patent owner would have enjoyed were it not for the activities of the infringer. If the court should

find that the infringement was willful and deliberate, the defendant may be ordered to pay the plaintiff's reasonable attorney's fees, but this award is relatively rare. Should the court determine that the infringement has been flagrant and without any justification, it may award up to three times the damages actually found; however, such an award is rare.

If the accused infringer wins the suit, he or she normally recovers little more than court costs. In some exceptional cases, particularly when the court considers the plaintiff's legal action to be an abuse of the judicial process, the plaintiff may also be ordered to pay the defendant's reasonable attorney's fees.

5.4 Settling Suit

Because of the expensive character of patent litigation, it is common for patent infringement suits to be settled before they come to trial. As discovery proceeds during the initial phases of patent infringement litigation, it is not uncommon for both sides to ascertain weaknesses in their positions and to note the desirability of settlement.

One point at which settlements are commonly effected is before a suit has been filed. In this situation, the patent owner confronts the infringer with the facts of the infringement, and a settlement is negotiated. Another time when settlement is commonly achieved is after there has been some initial discovery. By this time the positions of the parties have been clarified, and each party can begin to evaluate the other's strengths and weaknesses. A third point is after each side is ready to go trial. At this point, both sides clearly know the strengths and weaknesses of their cases, and they each may be anxious to eliminate the cost of an actual courtroom confrontation.

5.5 Declaratory Judgment Actions

In the event that one is accused of infringing another's patent or one's business is threatened by a possible suit for alleged patent infringement, one may bring a suit in federal court to have the threatening patent declared invalid or not infringed. The Federal Declaratory Judgment Act, passed in 1934, enables an alleged infringer to seize the initiative in this way and become a plaintiff, rather than a defendant, in a patent suit.

For a declaratory judgment action to succeed, the patent owner must actually have threatened to sue either the allegedly infringing manufacturer or its customers. An actual controversy must exist between the parties.

What the alleged infringer does in filing a declaratory judgment action is to seek a ruling from a federal court that the accused is not infringing a patent and/or that the patent is invalid. Filing such an action forces a patent owner who is threatening an alleged infringer or its customers to "put up or shut up." Accused infringers sometimes file such actions when it has become clear that they are going to be sued by a patent owner, for the first to file is usually able to select which available forum is most convenient for its use and most inconvenient and expensive for its opponent.

5.6 Failure to Sue Infringers

The laws of our country do not require patent owners to take any positive action whatsoever to enforce their patents. Many companies obtain patents and hold them defensively to preclude others from future use of certain inventions and as safeguards against the possibility of competitors patenting the same inventions. This practice is harmful to no one, because patents do not take from the public anything that was already in the public domain but rather expedite the disclosure to the public of new inventions that will come into the public domain when the patents expire. For many companies, the principal value of their patents lies in the defensive uses they make of them.

5.7 Infringement by Government

If the federal government infringes your patent, the only action you may take is to sue for monetary damage. No injunctive relief is available. The same holds true if government contractors, operating with the consent of the federal government, use your invention to carry out the work of the federal government.

If state or local governments infringe your patent in carrying out a public service, much the same situation results. No injunctive relief will be available, but recovery of monetary damage for unauthorized use probably can be had.

5.8 Alternative Resolution of Patent Disputes

Because patent infringement litigation can sometimes cost each party millions of dollars from the filing of a complaint to the reading of the verdict, there has been an enormous interest in finding a better way that ensures greater expertise in the trier of fact, more confidentiality regarding information that must be disclosed, lower cost, and a faster pace.

Alternative dispute resolution (ADR) comes in many flavors. It may involve arbitration, mediation, the provision of an early evaluation by a neutral party, summary jury trials, and a variety of other approaches. Depending on the approach one selects, it may be faster. On the other hand, with some federal courts now running “rocket dockets” that force cases to trial within six to nine months after the filing of a complaint, it may not be faster.

Arbitration tends to be expensive, but it may eliminate the need for extensive discovery inasmuch as witnesses can be brought before the arbitrator by subpoena, if necessary. Mediation tends to be faster than a trial if both sides display a willingness to compromise. A feature of mediation is that it can be taken up at substantially any stage of the proceeding once the parties are willing to compromise and work toward settlement.

An approach that some have found worthwhile is to utilize a neutral party to provide an early evaluation. The evaluator is an experienced practitioner acceptable to both sides and selected for his or her expertise in technology and in patent trials. He or she reviews the written submissions of the parties, conducts meetings with the litigants, and then renders a nonbinding report that evaluates the strengths and weaknesses of each side. This may assist the parties in viewing their positions more realistically so a settlement can be reached.

An advantage of ADR is that people picked to hear the matter can be individuals who have appropriate expertise, thereby eliminating much of the need to educate the fact finder who is encountered at trial. Also, ADR can make it easier to maintain confidentiality than may be possible in the courtroom.

5.9 Interferences

An *interference* is a complex contest between applicants to determine who will receive patent rights. An interference may be encountered after your patent issues because the applicant in a pending application copies one or more of the claims from your patent or during the prosecution of an application because two applications owned by different entities contain conflicting claims. Since an interference can arise after a patent has issued, one must not destroy one’s records of early-invention-related activities simply because a patent has issued.

Each party to an interference must submit evidence of facts that prove when the invention was made. If a party submits no such evidence, that party’s earliest date is restricted to the date his or her application was filed. The question of priority is determined by a board of three administrative judges based on evidence submitted.

Priority of invention will normally be awarded to the first inventor to reduce an invention to practice. This is because the act of invention is deemed not to have been

completed until an invention has been conceived and reduced to practice. An exception to the first-to-reduce-to-practice rule arises when the first to conceive has exercised reasonable diligence in reducing the invention to practice, but his diligent efforts did not result in his becoming the first to complete a reduction to practice. When this exception applies, the period of diligence of the first to conceive must extend from a time just before the second to conceive began its activities through the time of reduction to practice by the first to conceive.

Most foreign countries have no equivalent to the U.S. notion of interference because they operate on a first-to-file basis rather than on a first-to-invent basis. There has long been a debate as to whether the opportunity to participate in an interference is beneficial to U.S. applicants, for a first-to-file system would eliminate the need for these costly and complex contests. Few can afford these contests.

6 PATENT PROTECTIONS AVAILABLE ABROAD

U.S. patents provide no protection abroad and can be asserted against a person or corporate entity outside the United States only if that person or entity engages in infringing activity within the geographical borders of the United States. This section briefly outlines some of the factors one should consider if patent protection outside the United States is desired.

6.1 Canadian Filing

Many U.S. inventors file in Canada. Filing an application in Canada tends to be somewhat less expensive than filing in other countries outside the United States. With the exception of the stringently enforced unity requirement, which necessitates that all the claims in an application strictly define a single inventive concept, Canadian patent practice essentially parallels that of the United States. If one has success in prosecuting an application in the United States, it is not unusual for the Canadian Intellectual Property Office to agree to allow claims of substantially the same scope as those allowed in the United States.

6.2 Foreign Filing in Other Countries

Obtaining foreign patent protection on a country-by-country basis in countries other than Canada, particularly in non-English-speaking countries, has long been an expensive undertaking. In many foreign countries, local agents or attorneys must be employed. The requirements of the laws of each country must be met. Some countries exempt large areas of subject matter, such as pharmaceuticals, from what may be patented.

Filing abroad often necessitates that one provide a certified copy of the U.S. case for filing in each foreign country selected. Translations are needed in most non-English-speaking countries. In such countries as Japan, even the retyping of a patent application to put it in proper form can be costly.

With the exception of a few English-speaking countries, it is not at all uncommon for the cost of filing an application in a single foreign country to equal, if not substantially exceed, the costs that have been incurred in filing the original U.S. application. These seemingly unreasonably high costs prevail even though the U.S. application from which a foreign application is prepared already provides a basic draft of the essential elements of the foreign case.

6.3 Annual Maintenance Taxes and Working Requirements

In many foreign countries, annual fees must be paid to maintain the active status of a patent. Some countries require annual maintenance fee payments even during the time that the application remains pending. In some countries, the fees escalate each year on the theory that the

invention must be worth more as it is more extensively put into practice. These annual maintenance fees not only benefit foreign economies but also become so overwhelming in magnitude as to cause many patent owners to dedicate their foreign invention rights to the public. Maintaining patents in force in several foreign countries is often unjustifiably expensive.

In many foreign countries, there are requirements that an invention be “worked” or practiced within these countries if patents within these countries are to remain active. Licensing of a citizen or business entity domesticated within the country to practice an invention satisfies the working requirement in some countries.

6.4 Filing under International Convention

If applications are filed abroad within one year of the filing date of an earlier filed U.S. case, the benefit of the filing date of the earlier filed U.S. case usually can be attributed to the foreign applications. Filing within one year of the filing date of a U.S. case is known as filing under international convention. The convention referred to is the Paris Convention, which has been ratified by the United States and by most other major countries.

Most foreign countries do not provide the one-year grace period afforded by U.S. statute to file an application. Instead, most foreign countries require that an invention be “absolutely novel” at the time of filing of a patent application in these countries. If the U.S. application has been filed prior to any public disclosure of an invention, the absolute novelty requirements of most foreign countries can be met by filing applications in these countries under international convention, whereby the effective filing date of the foreign cases is the same as that of the U.S. case.

6.5 Filing on Country-by-Country Basis

If one decides to file abroad, one approach is to file separate applications in each selected country. Many U.S. patent attorneys have associates in foreign countries with whom they work in pursuing patent protections abroad. It is customary for the U.S. attorney to advise a foreign associate about how he or she believes the prosecution of an application should be handled but to leave final decisions to the expertise of the foreign associate.

6.6 Patent Cooperation Treaty

Since June 1978, U.S. applicants have been able to file an application in the USPTO in accordance with the terms of the Patent Cooperation Treaty (PCT), which has been ratified by the United States and by the vast majority of developed countries. PCT member countries include such major countries as Australia, Austria, Belgium, Brazil, Canada, China, Denmark, Finland, France, Germany, Hungary, Japan, Mexico, the Netherlands, Norway, Russia, Sweden, Switzerland, the United Kingdom, and the United States. In filing a PCT case, a U.S. applicant can designate the application for eventual filing in the national offices of such other countries as have ratified the treaty.

One advantage of PCT filing is that applicants are afforded an additional eight months beyond the one-year period they would otherwise have had under the Paris Convention to decide whether they want to complete filings in the countries they have designated. Under the Patent Cooperation Treaty, applicants have 20 months from the filing date of their U.S. application to make the final foreign filing decision.

Another advantage of PCT filing is that it can be carried out literally at the last minute of the one-year convention period measured from the date of filing of a U.S. application. Thus, in situations where a decision to file abroad to effect filings has been postponed until it is

impractical, if not impossible, to effect filings of separate applications in individual countries, a single PCT case can be filed on a timely basis in the USPTO designating the desired countries.

Still another feature of PCT filing is that, by the time the applicant must decide on whether to complete filings in designated countries, he or she has the benefit of the preliminary search report (a first office action) on which to base his or her decision. If the applicant had elected instead to file applications on a country-by-country basis under international convention, it is possible that he or she might not have received a first office action from the USPTO within the one year permitted for filing under international convention.

6.7 European Patent Convention

Another option available to U.S. citizens since June 1978 is to file a single patent application to obtain protection in one or more of the countries of Europe, most of which are parties to the so-called *European Patent Convention (EPC)*. Two routes are available to U.S. citizens to effect EPC filing. One is to act directly through a European patent agent or attorney. The other is to use PCT filing through the USPTO and to designate EPC filing as a *selected country*.

A European Patent Office (EPO) has been set up in Munich, Germany. Before applications are examined by the EPO in Munich, a receiving section located at The Hague inspects newly filed applications for form. A novelty search report on the state of the art is provided by the International Patent Institute at The Hague. Within 18 months of filing, The Hague will publish an application to seek views on patentability from interested parties. Once publication has been made and the examination fee paid by the applicant, examination moves to Munich, where a determination is made of patentability and prosecution is carried out with the applicant responding to objections received from the examiner. The EPO decides whether a patent will issue, after which time a copy of the patent application is transferred to the individual patent offices of the countries designated by the applicant. The effect of EPC filing is that, while only a single initial application need be filed and prosecuted, in the end, separate and distinct patents issue in the designated countries. Any resulting patents have terms of 20 years measured from the effective date of filing of the original application.

6.8 Advantages and Disadvantages of International Filing

An advantage of both PCT and EPC filing is that the required applications can be prepared in exactly the same format. Their form and content will be accepted in all countries that have adhered to the EPC and/or PCT programs. Therefore, the expense of producing applications in several different formats and in different languages is eliminated. The fact that both PCT and EPC applications can, in their initial stages, be prepared and prosecuted in the English language is another important advantage for U.S. citizens.

A principal disadvantage of both of these types of international patent filings is their cost. Before savings over the country-by-country approach are achieved, filing must be anticipated in several countries, perhaps as many as four to six, depending on which countries are selected. A disadvantage of EPC filing is that a single examination takes place for all the designated countries, and patent protection in all these countries is determined through this single examination procedure.

CHAPTER 27

ONLINE INFORMATION RESOURCES FOR MECHANICAL ENGINEERS

Robert N. Schwarzwald Jr.
Stanford University
Stanford, California

1 BACKGROUND AND DEFINITIONS	805	5 GUIDE TO MECHANICAL ENGINEERING RESOURCES	813
2 OVERVIEW OF ONLINE INFORMATION	809	5.1 Database Services	813
3 ACCESS OPTIONS FOR MECHANICAL ENGINEERS	809	5.2 Document Delivery Options	817
3.1 Access Options through Corporations and Universities	809	6 MANAGING YOUR ONLINE LITERATURE	819
3.2 Independent Access to Online Information	811	7 FUTURE OF ONLINE ACCESS	819
4 ONLINE SECURITY	811	8 OPTIONS FOR USING ONLINE INFORMATION	821
		REFERENCES	822

1 BACKGROUND AND DEFINITIONS

Rapid and convenient access to information has become a given in the twenty-first century. Information resources once limited to a few corporate or research libraries are now readily available. In this age of information the question is no longer, “How can I find that?” but it has become, “How can I keep up with the flood of information in my field?”

While there is a wealth of information available online, not everyone is well versed on where to look. Engineering classes still rely too heavily on course packs and limit students’ exploration of the engineering literature. It has been repeatedly observed by information scientists and librarians that engineers make less frequent use of the technical literature than do scientists or other similar professionals.^{1–3} Possible explanations for this include the esoteric nature of many technical publication series (technical papers, standards, government publications, etc.) and the difficulty many professional engineers have in gaining access to these publications. By making the less available technical literature available to the desktop, the Internet has by-passed some of the obstacles to access and created an online option for engineers. This explosion of information, complete with electronic books and journals, databases, Internet services, and digital archives, has not necessarily made it easier to obtain the information you need. When confronted with the thousands of results from an Internet search, how do you respond? How do you pick the right options? Is the information you find trustworthy? And, what if the information you need is not among those results? In addition to the bewildering array of options, the once-bucolic Internet has become somewhat perilous. Beyond the inconvenience of SPAM and pop-up windows are the very real hazards of email viruses and malware.

The Internet has matured as an information resource, as have the options, risks, and opportunities open to you as an engineer. While the Internet can become your desktop information center, it can also be a source of frustration. Conducting Google searches of the Internet may be perfect for many inquiries, but you may find it an undependable strategy for obtaining technical information. However, there are a variety of services available that will provide you with rapid access to professional information. Your access to these collections and tools will depend upon your work environment. Within universities and large corporations you should be able to use online tools and collections licensed for use in your organization. If you work independently or in a small company, you may need to privately contract to obtain access to these resources. Once connected, access to full-text versions of books, journals, standards, and patents will allow you to access the world's technical literature from your desktop.

In this chapter I will provide an overview of how you can navigate the maze of professional information resources available through the Internet while minimizing the frustrations and pitfalls inherent with the Web. I will discuss approaches for gaining access to suites of online services and content for mechanical engineers in a variety of job environments. I will also provide a listing of selected online resources for mechanical engineers.

Engineering information turned a corner in the late 1990s. As technical literature, standards, and engineering data blossomed on the Web, libraries and information service providers flocked to the new medium. Today, the speed of obtaining online information and the ease of embedding that information in proposals and reports has changed the way the profession works. If you have discovered online information, you will know how much time and effort these new services save. If you have not, you may find that the opportunities involved are well worth changing the way you obtain information.

The following terms are related to the discussion of the Internet and online information resources:

Adware. Advertising software that is downloaded onto your computer from the network that either continues to advertise a service or product on your computer or functions to alert a third party of your activities. Some adware serves a useful function such as alerting you to new versions of a product; and some adware is clearly malicious, operating without your permission and against your wishes.

Agent (Information Agent). A software device that filters information before it reaches the user or locates and sends information to the user.

Cloud Computing. In cloud computing, storage or computation services are offered as a remote service through an interconnected network of hardware and software. The same concept is often described as “Software as a Service” (SaaS), “Storage as a Service” (STaaS), or “Desktop as a Service” (DaaS).

Client. A software application mounted on your computer that extracts some service from a server somewhere else on the network. This relationship is often referred to as a *client-server* application.

Cookies. Web cookies are files created during a Web session that retain information about your identity and preferences. Cookies are used to help personalize a Web session or retain information required during a complex transaction. Cookies are essential to many online commerce applications and are typically deleted at the end of a session. Persistent cookies are retained on your hard drive.

Database. A computer-based search and retrieval system that allows a user to retrieve and display information based upon a series of command protocols.

Datafile. A database containing numerical, chemical, or physical data.

Data Management Plan. A plan describing the generation, management, and preservation of data generated during a research project, typically submitted as part of a grant proposal. In the early 2010s a number of federal and private agencies funding engineering and scientific research began to require data management plans that detailed the steps researchers would take to make their data accessible to others.

Descriptor. A term used by a database producer to index a database record. Descriptors provide a consistent tag that can be used to retrieve all items relating to a given topic.

Digital. A digital item is one that is available as a computer file and can be accessed online.

Downloading. Refers to the transfer of electronic data from a server to another computer, usually a personal computer.

DSL. A digital subscriber line, or DSL, is a high-speed, high-capacity connection to the Internet through telephone lines.

DSpace. An open-source initiative begun by MIT Libraries and Hewlett-Packard that allows institutions to create digital repositories. DSpace initiatives seek to preserve free access to scholarly information.

DVD. Digital video discs, or DVDs, are optical storage discs. A single-sided DVD is capable of storing between 1.46 and 8.54 GB of data.

FAQ. Often a list of “frequently asked questions” appears on a website. These questions and answers are intended to handle routine inquiries. FAQ lists are a good place to go if you are having problems using a website.

Federated (or Meta) Searching. Allows the user to simultaneously search more than one database and retrieve a combined set of results.

Fedora. A collection of *open-source* software including the core software used for a number of large *institutional repositories*. Much of the active development in this area is now carried out by developer communities for “Hydra” and “Islandora.”

File Server. A host machine that stores and provides access to files. Remote users may use File Transfer Protocol (FTP) to obtain these files.

Firewall. A security system designed to keep unauthorized users out of a computer or computer network. Network firewalls are often a combination of hardware and software while personal firewalls typically consist of software alone.

GUI. (Pronounced Goey) A graphical user interface, or GUI, is a system like the Web that allows users to view and use graphics, as opposed to a text-only interface.

Host. A network computer that contains resources that are shared by other members of the network.

HTML. HyperText Markup Language is a system of computer language tags that are used to create Web pages.

Identity Theft. Involves the misrepresentation of a person for financial gain using personal data stolen from computer files or other sources. Often identity theft is the motivation of malware developers.

Institutional Repository (IR). An archive of materials maintained by a university or academic institution with the purpose of preserving intellectual content or providing access to materials not restricted by copyright.

Internet. A collection of interconnected networks that speak the Internet Protocol (IP) and related protocols. The Internet provides a variety of services, including email, file transfers, streaming video, and the Web.

Intranet. An internal Internet, a collection of interconnected computers or networks open to members of an organization but closed to external use through a variety of security protocols.

Listserv. An email list devoted to a specific topic. Any message posted to a Listserv is forwarded to all of its members.

Malware. A broad term referring to any undesirable software, typically software that is installed on a computer without the owner's knowledge or permission. Malware consists of spyware, viruses, and some adware.

MOOC. A massive open online course is a form of distance education wherein an online course is opened to a very large audience. Courses are at the university level but are not offered for university credit.

Online. An online, or networked-based, resource is available to remote users through the Internet.

Open Source. A movement that developed in response to price escalation from large software developers. While terms of use vary significantly, most open-source software is free for anyone to use or modify as they wish.

OpenURL. A protocol and standard (ANSI/NISO Z39.88-2003) that allows the user to link directly from a database citation to the digital version of an article.

Operating System (OS). A set of software applications that manage the basic functions of a computer or mobile device.

PDF. Portable Document Format is a computer platform independent electronic file format developed by Adobe Systems, Inc. PDF documents have become a popular standard for reproducing documents on the Web since the format preserves the pagination and appearance of the original document.

Remote login. The process of accessing a host mounted on another network.

Server. A term used to refer to (1) software that allows a computer to offer a service to another computer (i.e., client) or (2) the computer upon which the server software runs.

Shibboleth. An open-source initiative to create a shared protocol for shared access, network security, and rights management. While in an early stage of development at the time of this writing, Shibboleth could become a major access and security protocol in the next few years.

SPAM. Undesired email, typically containing appeals to purchase questionable products, visit questionable websites, or engage in questionable activities.

Spyware. A term given to invasive or malicious software that monitors the activities on a computer and reports back to a third party.

URL. A universal resource locator, or URL, is an electronic address on the Web.

Virus. A malicious computer program that replicates itself and spreads through the Internet. Nearly all viruses cause some harm to the files or software on a computer, although some viruses are used to hijack a computer to send SPAM or for purposes of identity theft.

2 OVERVIEW OF ONLINE INFORMATION

There is a wealth of engineering information available through the Internet. With a basic awareness of the range of information content and services open to you, a computer with Internet access, and a major credit card, you can have greater access to the current engineering literature than at any library in the world. Unlike much of the material available through the Web, these resources are not garbage and they are not free. The advantages of having access to these resources anytime or anywhere you wish are truly revolutionary. For the ill prepared the risks of wasting time and money are significant.

The focus of this chapter is to make you a savvy information consumer and to introduce you to resources and strategies to make you an effective and efficient consumer of these services. There are some excellent services in the Internet marketplace that will make your task far easier. Unfortunately, like any marketplace, the Internet also harbors thieves and hustlers. Before doing business on the Web, please consider the points raised in Section 4. Even if you feel safe within the firewall of a corporate intranet, you still could fall prey to Internet scams, viruses, or malware. Your vigilance is your and your organization's first line of defense.

3 ACCESS OPTIONS FOR MECHANICAL ENGINEERS

How you access online engineering information will largely be a function of the nature of your employment. If you are working in a corporation or large company or if you are employed in a college or university, you should have access to a large collection of online collections through your corporate or university library. If you are working independently or are employed by a small- or medium-sized company that does not have a library, then you will want to contract for information access independently. While the majority of information companies prefer to work with large organizations and do not provide options for individuals to purchase access, I provide a number of options in Sections 5.1 and 5.2 for free or fee-based information access.

3.1 Access Options through Corporations and Universities

In most corporate and academic environments, your library will have subscribed to a collection of information resources and you will be able to use these resources without charge. The easiest way to get familiar with these collections is to contact your library and ask them what resources exist for mechanical engineers. You may also want to inquire about business resources since the engineering trade literature is sometimes covered by business databases. In addition to books and journals, most major corporate and university libraries provide a wealth of additional material of relevance to the mechanical engineer. Collections of industrial standards are common in

the libraries of large manufacturing companies and academic libraries serving large engineering schools. While patent collections are less common, the patent literature often provides one of the few avenues to explore proprietary research. The technical reports literature can also be quite valuable; however, technical reports can be very difficult to obtain.

If you have access to a large library, most of these materials will be accessible through the website provided by your library or information center. Books, journals, and conference proceedings can be found on the main online catalog, although publisher portals play a major role in providing specialized access in some areas of specialization. For finding individual articles in journals or conference proceedings, however, you will want to use an article database.

In summary, for most questions you will find relevant information in four different ways:

- *Online Catalog.* The online catalog allows you to search for books, journals, conference proceedings, and technical reports held by your institution's library. It will not identify articles within journals or conference proceedings or an item that is not held by that library. Library online catalogs allow searching by keyword, subject term, author name, or title. When an item is available as a digital file, the online catalog will typically link directly to the book or journal through an OpenURL service. Where multiple versions of the item are available, the OpenURL service will provide a list of options.

- *Online Books and Journals.* The publishers of engineering books and journals have moved strongly toward making their publications available through the Web. Large to medium corporate and academic libraries will have significant collections of online engineering literature that you can access directly from your lab, home, or office. While you may be able to find these items through the online catalog, many publishers provide their own platforms that provide better and more complete access to their publications. This is especially the case with eBooks, where the records for individual eBooks tend to show up in online catalogs sometime after they appear on a publisher's platform. Unfortunately, the publishers of trade literature have a more spotty record regarding online access.

- *Engineering Databases.* To locate articles within engineering journals or conference proceedings, you need to search your topic in a topic-specific database. I have provided details on some of the major article databases for mechanical engineering below. At this time a few libraries have implemented "federated searching," or "meta-searching" technologies that allow you to search multiple databases without selecting the individual database for your field. In most cases, you will need to know the correct database to use for a given topic. These databases will allow you to locate citations and abstracts of articles that may be of interest. Once you have obtained these citations, you will need to locate the articles. If an OpenURL resolver is available, you can link directly to the digital copy of the article from the database citation. If your library has not implemented OpenURL, you will need to copy the information in the citations you obtain from the database and then go into the online catalog to determine if your library holds the items. If your library does not own the desired items, you may request copies through your library's document fulfillment service or Inter-Library Loan service. These services often involve an expense and a delay of days or weeks.

- *Ask a Librarian.* Most corporate or academic engineering libraries will have staffs that are well trained in the engineering literature. Often these individuals have advanced degrees in engineering or work experience in related industries. A good engineering librarian can help you identify where you might find information related to your area of concern and can work with colleagues across the country to locate hard-to-find materials. These professionals are valuable allies in your search for information.

3.2 Independent Access to Online Information

It is possible that if you are not affiliated with a university, you may still benefit from their engineering collections. Public universities retain the right to allow the general public to use their print and digital collections on site. If you visit the library of a large public university, you can typically use these digital collections from a public terminal in the library. These privileges do not extend to remote users unless they are directly affiliated with the institution. Corporate libraries never offer this option and private universities have varying policies.

If you cannot gain access to these resources through a public university or if you prefer the efficiency and convenience of using these materials from your office or home, you have some options for obtaining online access to the mechanical engineering literature on a fee-per-use or subscription basis. In the late 1990s and early 2000s a number of companies began to sell this type of access, believing the demand would be strong. A less than robust demand and the economic turbulence of the last decade have diminished the options for individuals looking to subscribe to an information resource that provides the full range of coverage available through a large academic or corporate library. However, a variety of options for online information access are provided in the following sections of this chapter.

4 ONLINE SECURITY

The Internet has become a major avenue to find and acquire technical information. Content providers and information services have rushed to establish businesses on the Web as more and more engineers, academicians, and business people have discovered the efficiencies of working online. This new interest in the Web has also attracted a criminal element intent upon destruction and fraud. While these miscreants are not a reason to avoid using the Internet, you will want to exercise care in your online dealings.

If you are using the Internet through your company or university intranet, the people who administer your system will have provided several layers of protection. These internal networks typically screen email and file attachments for viruses and provide firewall protection against hackers who attempt to enter the system without permission. This does not mean that you should not be vigilant, but it offers some level of protection if you are not. If you are accessing the Internet from your personal computer, you should consider subscribing to a virus protection service as well as a firewall service. Since the people responsible for these threats are constantly developing new software, you need to subscribe to a service that is constantly updating their protective software. A virus protection service will recognize and stop a virus when it attempts to infect your computer.

A firewall service will prevent hackers from gaining access to your computer. Firewalls are especially important if you are accessing the Internet through a cable modem or DSL (digital subscriber line). Hardware firewalls are often a built-in feature of wireless access points. Using hardware such as a Cisco wireless access point or an AirPort Extreme router/Wi-Fi base, you gain a large measure of protection ... as long as you implement the product as a secure access point. In addition, some of the most popular antivirus software packages also have options that include software firewalls.

While a great deal has been written about computer viruses, it is useful to review a bit of it here. Computer viruses are self-replicating software packages that install themselves on a computer and send copies to other computers on a network. Sometimes viruses are benign and do little harm; in other cases they can destroy content, raid private files, or turn over control of your computer to a third party. The fact that new viruses are being created on a continuing basis makes it important to subscribe to a service that is constantly updating their protection.

The two best-known providers of these services are McAfee (<http://www.mcafee.com/us/>) and Symantec (<http://www.symantec.com>). Firewall and virus protection services will automatically recognize threats as they arise and do not require much effort to use. Many viruses are carried as email attachments and all email users are urged to take the following steps:

- Do not open attachments of email messages that seem generic, impersonal, or out of character with the sender. (Email viruses often send messages from an infected machine, so knowing the sender is no protection at all.)
- Do not open zipped attachments, attachments with an .exe file extension, or attachments with unusual file extensions unless you are expecting these formats.
- Be cautious if you get a large number of messages with similar subject lines in a brief time period; they may contain viruses.
- If you are using a Microsoft Windows operating system, visit the Microsoft website (<http://windows.microsoft.com/en-US/windows/downloads>) and look for critical security patches on a regular basis.
- Do not reply to SPAM ever! (Some SPAM asks you to send a reply if you want to be taken off of their emailing list. This is a trick! Don't do it!)

A new threat to online users has developed which is currently less well understood but is as potentially dangerous as viruses. Spyware is a term applied to software that resides on your computer and reports back on your activities. Some spyware is relatively harmless and monitors your use of a site so that the provider can personalize its service. Some spyware observes your usage of other sites on the Internet for a variety of reasons. And some spyware records your keystrokes in an attempt to steal credit card numbers and passwords. You often unwittingly download spyware when visiting websites or when you download a piece of software from the Internet. At best, spyware slows down your PC, sometimes to the point that it appears to be malfunctioning. At worst, spyware is a major identity theft risk. To avoid problems with spyware, you should:

- Read the terms and conditions of any software you download very carefully. If the provider mentions monitoring your computer or your usage, do not accept the terms.
- Never accept downloads of software that appear in pop-up windows when you are visiting an unknown website, unless those downloads are from a well-known company (Microsoft, Adobe, etc.).
- Set the preferences on your Web browser. Depending on your browser, you should be able to require permission to install an application or add-on, block pop-up windows, restrict the ways JavaScript can be enabled, etc.
- Install and use, on a regular basis, a program to find and eliminate spyware from your computer. A couple of options are *Ad-Aware* (<http://www.lavasoft.com>) and *Spybot* (<http://www.safer-networking.org/>).

At this time the majority of viruses and spyware are written against the Windows Operating System. This speaks more to the popularity of Windows and PCs in office and home environments. Apple products are not exempt from these attacks but are less frequently targeted. The fact that mobile devices have not experienced the same degree of problems with malware is more a function of their recency on the market. For now, the odds of contracting a virus on a mobile device are low. However, these devices rely upon an operating system (OS) and are, therefore, potentially vulnerable.

5 GUIDE TO MECHANICAL ENGINEERING RESOURCES

In my chapter in the last edition of this handbook I presented a mixture of fee-based database services as well as a number of free Web-based resources for mechanical engineers. The online landscape continues to evolve. In the early to mid-2000s a number of companies involved with information services anticipated the decline of libraries as more and more individuals established individual accounts for access to technical information. The cable television business was seen as a model for how people would pay for information and news. As a result, most of the large information service companies provided individual access options and began to market directly to end users. The wholesale development of Web-based access systems was seen as facilitating a vast new business to deliver information to the desktop.

Everything seemed aligned in favor of this new business model, except that the market never developed. There has never been a mass market for this type of information and most of the people who need it require a certain degree of intervention in locating the literature. Unfortunately, the information industry has swung in the opposite direction in the last five years, eliminating many of the options they had created for individual access. While I think that, in the future, new services will be created to address this niche market, for the time being it may be more difficult for a consulting engineer or an engineer working in a small firm to get full access to the engineering literature. That said, there are approaches that can be used to search for technical information and obtain copies of articles, standards, and reports.

There is a great deal of free access to information available on the Web; one just needs to be cautious about the validity, accuracy, and currency of what is retrieved. The U.S. federal government provides access to a wealth of free information and has embraced the Web as an effective communications vehicle. The government Web service most directly applicable to engineering is the U.S. Patent and Trademark Office's (USPTO's) free patent system, which is described below. Google Scholar provides an excellent index of scholarly publications and, while the coverage is not as deep as some specialized databases, strong overall indexing at zero cost.

The reduction in online resources available to the end user does not suggest that the Web has diminished as a vehicle for online engineering information. The speed, convenience, and popularity of the Web have made it the primary means through which to access online databases, electronic publications, and a wide range of information services. Direct dialups to database networks through computer modems is largely a thing of the past, as are databases on CD-ROM. This move to Web access has expanded access and allowed vendors to use more attractive and effective interfaces and to introduce electronic information access to a whole new audience.

5.1 Database Services

The information industry has grown considerably more complex in the last two decades. By the early 1990s the companies that produced paper indexes to the engineering literature had developed computer-based citation databases. At that time, their only way of marketing these databases was through large commercial networks—like STN or Dialog—which provided dial-up access to academic and corporate clients. By the mid-1990s the database producers had found that they could market their databases directly to large institutions and to individuals in the form of CD-ROMs. A few companies had begun to use the Web to provide fee-based database searching by the mid-1990s, a business practice at first widely criticized and then almost universally accepted. By the end of the 1990s, the large database networks were seeing defections as database producers began to perceive a more lucrative direct market through the Web. In the late 1990s and early 2000s database producers marketed heavily to end users, primarily through the Web. By the 2010s a weak end-user market and the advance of enterprise tools like federated searching and OpenURL resolvers had focused the market for these services on large academic and corporate clients.

How does it translate to you on a personal level? If you are employed by a large- to medium-sized organization, corporate or academic, you may find that your employer has contracted directly with specific database producers to provide anyone employed there with database access. This is typically done through the Web. The database producer recognizes you as a member of a subscribing organization and grants you access. An evolving set of tools makes your discovery of new information, your access to relevant documents, and your management of digital files far easier than at any point in the past. As mentioned, if you are not affiliated with one of these institutions, your access to the engineering literature is far more limited.

In the text below I have provided lists of the major sources of online information and a sampling of databases relevant to mechanical engineering. This will provide you with a basic guide as to which database network might best meet your needs.

Database Resources for Mechanical Engineers

This section provides an overview of the types of resources available to you that can be used to locate engineering information. In most cases, you will use a database to search for citations to relevant information. In some cases that database will provide direct access to the item referenced by the citation; in other cases, you will need to use another service to obtain a copy of the item. (Document delivery options are discussed in Section 5.2.) When selecting a database you need to consider the following aspects:

- The subject coverage of the database—does it match your search topic?
- The type of materials covered by the database—if you want numerical, physical property, or chemical data values you should not be searching an index of the journal literature; you need a datafile.
- The coverage of the database—if you need international coverage, you cannot just search databases that index information from the United States.
- The availability of the materials indexed in the database—while the National Technical Information Service (NTIS) is an excellent source of U.S. technical reports, it will not help you if you cannot access these documents in time to meet your deadline.

When searching databases most beginners are either too general or too specific. If your search query is too broad, you will retrieve a large set of records, most of which will have little relevance to your desired topic. If your search is too specific, you will retrieve nothing or a small subset of what the database contains on your subject. In the first case, you will waste time and money; in the second, you will miss vital information. The best way to approach searching is to start with a fairly broad keyword or phrase that describes your topic. If the results seem too numerous or miss the mark, refine the search by adding in other keywords or concepts. By continuing to refine the search in this manner, you will arrive at a manageable set of results.

Another approach to searching is to find a few records that match your interest, display them, and determine which subject headings the database uses to describe the topic and then use those subject headings to conduct a follow-up search. Most databases use subject headings, sometimes called “descriptors,” to consistently index topics. A good database will have a thesaurus with a listing and explanation of its subject headings. If you are able to use subject headings in your searches, you will improve both the precision and comprehensiveness of your searching.

In most searches the balance between the precision and comprehensiveness is a major issue. Most people intend to perform precision searches in order to get a few good references on a topic. If you attempt to expand that search, you can reduce the number of terms in your search, include alternative terms relating to the same idea (e.g., engines, motors, powertrains), or use a broader term for the topic (e.g., *welding* in place of *arc welding*). However, by being

more inclusive, you may retrieve a very large set of results. If you need to winnow down a large set of results, I would try two alternatives. First, limit your search to information published in the last few years. While these may not be the best articles on your topic, they will be current and should reference any major articles written in the last several years. Another strategy is to include title words such as “review,” “summary,” or “progress” in your search. (You can typically limit searches of particular keywords to specific fields, such as title, subject, or abstract.) These words tend to appear in the titles of articles that review progress in a given field. These review articles, when available, are a great way to get an overview of a field of investigation. Also, if you are investigating a technology that is new to you, do not forget to consult a good technical encyclopedia or handbook. These resources often provide a good starting place for a new project.

Fee-Based Database Service

STN (Paid Service). STN International is a network of scientific and engineering databases run through centers in the United States, Japan, and Germany. STN is an excellent choice for individuals who want to establish individual accounts and offers a limited but attractive list of engineering databases. *STNEasy* is a service marketed to the end user and provides a low-cost option for establishing an account with a pay-as-you-go model for access. Information on *STNEasy* can be found at <http://www.cas.org/products/stn/easy>.

STNEasy provides a number of databases that will be highly relevant for mechanical engineers:

- **COMPENDEX** is one of the foremost indexes to the engineering literature and does an excellent job of covering the English language mechanical engineering literature. **COMPENDEX** covers materials published from 1970 to present and contains over 9 million records covering all aspects of engineering. The database has a consistently applied set of index terms, descriptors that help improve the precision of your searches.

- **WPINDEX** (Derwent World Patents Index) is an excellent index of the major patent-issuing bodies in the world. It contains over 21.7 million records and covers mechanical, electrical, and general technology patents from 1974 to present. While you can search U.S. patents for free (see Internet Resources below), this database enhances its database records with subject terms that make them easier to search and provide information linking related patents from various countries, and the EU, into patent families. This is critical information for anyone wishing to market a product outside of the United States and a huge advantage of this database.

- **METADEx** is a database that provides global coverage of the literature on metals and metallurgy. It indexes the literature from 1966 to the present and contains more than 2.3 million records. While **COMPENDEX** does address this discipline, **METADEx** provides far deeper coverage and contains related information in areas such as the physics of metals, mining, and materials properties that would be harder to find in **COMPENDEX**.

- **NTIS** is a database of primarily U.S. government technical reports. The database contains over 2.4 million records going back to 1964. **NTIS** is a poorly indexed database, resulting in a lot of odd results when you search. In addition, the reports indexed in **NTIS** are some of the most difficult materials to obtain. That said, the U.S. government conducts a great deal of engineering research, much of it is otherwise proprietary and nearly impossible to find documented. This literature is some of the most overlooked in engineering. For topics where you suspect governmental funding, **NTIS** is a database you should consult.

- **WELDSEARCH** is a specialized database covering the global literature on welding. It covers the literature from 1967 to the present and contains over 200,000 records. While far smaller than other resources mentioned here, it does address a specialized need.

Fee-Based Services Common in Academic or Corporate Environments

These services do not provide the option to establish an individual account; however, they are common in universities and corporations serving engineers or engineering students. To locate these services go to your library or information center website and search the name of the service or contact your engineering librarian to inquire.

Engineering Village (www.engineeringvillage.com/). Engineering Information, Inc. was purchased by publishing giant Elsevier in 1998. The “Engineering Village” was created before the acquisition as an attempt to provide a content and service destination for engineers looking for technical information. Elsevier has continued to refine the service, streamlining the search interface, adding features, and showing the number of times the item has been cited in Scopus. (See the section on Scopus below.) The Engineering Village provides access to a number of databases, including Compendex and NTIS, and will combine results from various databases when you conduct a search. Searching is faceted, allowing you to winnow results by year, language, descriptor, or a variety of other factors. The system is OpenURL compliant, so at most academic institutions you can link directly from the citation to the full text of the article. This service provides a nice suite of databases for the engineer along with a clean, easy-to-use user interface.

Knovel (<http://why.knovel.com>). Knovel, pronounced “Novel,” is a unique service directed at engineers and scientists. Knovel has developed a technology that adds functionality to engineering and scientific handbooks. Knovel’s interactive tables allow you to find data in a table, filter and export data within tables, sort the contents of tables, graph and interact with data from selected data from tables, and much more. The software even allows you to digitize the images of curves and graphs and export those data. Handbooks are a vital literature for mechanical engineers and Knovel has done a remarkable job of adding new functionality to this old standby.

Scopus. Scopus claims to be the world’s largest abstract and citation database of the world’s peer-reviewed literature. With 53 million records covering 21,912 journals, 5.5 million conference papers, and 421 book series, they make a strong case. Beyond offering a well-designed search system that is OpenURL compliant, Scopus provides a number of special features that make it extremely popular with users. The Citation Tracker feature identifies other papers in Scopus that cite an item of interest and allow you to link to the citing papers. You can establish alerts that will send you update lists of citations that match your search parameters. In addition, Scopus allows you to export records into a variety of bibliographic managers. (See Section 6.)

Free Services on the Web

Engnet (<http://engnetglobal.com/>). Engnet is a free directory of products and services related to engineering. The site is browseable by product category, brand name, and discipline. The discipline listing includes automotive, manufacturing, and mechanical. The directory includes hundreds of engineering product and service companies from around the world. You can filter results by country or region.

Nanohub (<http://nanohub.org/>). Nanohub is an online community and repository devoted to nanotechnology. The site is so diverse it can be difficult to describe adequately. Nanohub serves as a repository for presentations, papers, online course materials, data sets, software, and whatever else people feel is important to the study of nanotechnology. Members of the community provide commentary and discussion. The site serves as a community for almost a quarter million users annually and, if we are lucky, will serve as a model for other technical communities.

United States Patent and Trademark Office (USPTO) Patent Database (<http://www.uspto.gov/>). The USPTO provides free searching and display of granted patents and patent applications. While this service does not provide the indexing of the Derwent databases (above) and includes only U.S. patents, it is a free service. At the very least, it is a very good place to start your patent searches and a wonderful service of the USPTO. The patent literature is one of the few ways to search for proprietary research. However, as these materials are written by patent attorneys, and not engineers, searching is made more difficult by the use of nonstandard terminology. As mentioned, the Derwent patent database available on STN adds more common index terms to the patent records, making them easier to search.

TRAIL (<http://www.crl.edu/grn/trail>). The Technical Report Archive & Image Library (TRAIL) is an initiative to collect, digitize, and provide access to a collection of government technical reports. TRAIL has been more aggressive and more innovative than the U.S. government in providing easy access to technical reports on the Web. The emphasis is on older technical reports, which are typically extremely difficult to locate. While this is not a service many people use frequently, it is useful to know about should you ever require access to this often inaccessible engineering literature.

eFunda (<http://www.efunda.com/home.cfm>). eFunda, short for Engineering Fundamentals is a website designed to provide quick reference information for engineers. Their goal is to provide formulas and other engineering information with concise background information to advise engineers on when and how to use the information. eFunda provides some information for free (e.g., composition of AISI Type 201 steel) and charges \$10 per month for access to an online library of additional material (e.g., materials data sheets, engineering formulas with online calculators).

Google Scholar (<http://scholar.google.com/>). For years the common advice offered by information professions was to avoid Google when searching for engineering literature because search results might contain material that was out of date, lacking peer review, or simply incorrect. Google Scholar addresses that concern by limiting results to the scholarly literature. Its coverage is quite strong in all fields and covers English language and non-English language materials. A Google Scholar search will retrieve citations to the literature, an indication of the number of times that item has been cited, and often links to commercial sources for obtaining a copy. While Google Scholar may not replace the need for a dedicated engineering database, it is often a great place to begin a search of the technical literature.

5.2 Document Delivery Options

The primary motivation for many engineers when going online for technical information will be to obtain a copy of a published article, standard, or book. It might be as a follow-up from a database search or simply to find a work referenced by a colleague or cited in a paper. Most large corporations and university libraries provide OpenURL linkages between their database and journal subscriptions so that users can go directly from a database citation to the full-text documents. For more obscure documents (foreign standards, governmental technical reports, etc.) a librarian or information professional is often on staff to assist. For the engineer not affiliated with a large corporation or university, there are a variety of options for obtaining documents.

In obtaining engineering materials you should look to either the organization that published the work or a large document supply company. Going to the publisher is often the best

approach with industrial standards and societal publications. Standards are best supplied by the originating society since that group will be able to provide the most up-to-date version of the standard. As standards are frequently revised or superseded, it is critical to obtain a current version. In the case of other societal publications, it may be necessary to go to the originating society to get the full range, or most current, publications. The list below features both specialized societal publishers and mainstream document suppliers.

eStandards Store—ANSI (American National Standards Institute) (<http://webstore.ansi.org/>). The ANSI eStandards Store allows you to purchase copies of a wide range of U.S., foreign, and international standards. Among others, the site provides access to ANSI, ASTM, ASME, and ISO standards. In addition to purchasing single copies of standards, you can establish a site license for individual or groups of standards. A site license may be preferable for certain families of standards, such as ISO 9000, that may be referenced repeatedly by a group of engineers.

American Society of Mechanical Engineers (ASME) [<http://asmedl.org/> (Digital Library) <http://www.asme.org/lists/product?orderBy=Date> (Product Catalog)]. ASME provides several options for accessing its journal articles, conference papers, standards, eBooks and, of course, the Boiler and Pressure Vessel Code. The “Digital Library” requires a subscription and, while there are a variety of subscription options, the Digital Library seems to be marketed at libraries rather than individuals. A better option for the individual engineer is to use the “Product Catalog” to purchase individual documents or eBooks.

Society of Automotive Engineers (SAE) TechSelect (<http://www.saedigitallibrary.org/corporate/small-business/techselect/>). SAE Technical Papers are an important source for automotive engineering; however, they are surprisingly difficult to obtain due to SAE’s rather insular policies regarding publications. TechSelect provides a relatively low cost option for individuals or small organizations to obtain copies of roughly 100,000 technical papers dating back to 1906. As the SAE technical papers are poorly indexed by many of the large engineering databases, it is best to search for these materials through SAE.

Infotrieve (<http://www.infotrieve.com/document-delivery-service>). Among a variety of services, Infotrieve is a major provider of document delivery services with offices throughout North America and Europe. Infotrieve supplies around a million documents a year, most of them electronically with very fast turn-around time. Infotrieve is a good option for individuals but can also provide services to small companies or large corporations.

IngentaConnect (<http://www.ingentaconnect.com/>). IngentaConnect provides free access to search a database of over 5 million scholarly articles. The IngentaConnect business model provides free searching but makes money by selling access to the articles referenced by the database. It is included here, as opposed to the database section, because while IngentaConnect is not the best place to search for the engineering literature, it does provide a convenient way of purchasing individual articles. (Like Scopus, you can export IngentaConnect search results into a bibliographic management tool. See Section 6.)

British Library Document Supply Services (<http://www.bl.uk/articles>). If you need to find a copy of a conference paper for a regional society in Bulgaria, a 70-year-old technical standard from Japan, or some other obscure publication, you might consider the British Library’s Document Supply Services. The service is known and respected worldwide for their speed, accuracy, and access to the global scholarly literature.

6 MANAGING YOUR ONLINE LITERATURE

The advent of online citations and digital articles has allowed engineers to manage their personal libraries in new, more effective ways. The office fixture of rows of filing cabinets has given way to collections stored on personal computers. Personal libraries are now accessible on a PC, a flash drive, or the cloud. Beyond the convenience of access, the software available to manage these collections allows far greater functionality than the old paper files. As the amount of the engineering literature available digitally has increased, these tools have become more popular. A number of major database services now facilitate the downloading of material into a personal bibliographic management system through simple point-and-click interfaces.

Bibliographic management software has been available for decades; however, most of these packages have been expensive and limited in functionality. While there are commercial systems in common use, especially EndNote and ProCite, I would recommend you consider two open-source options: Zotero and Mendeley.

Zotero (<http://www.zotero.org/>). Zotero allows you to capture citations, PDFs, audio and video clips, Web pages ... pretty much whatever you can find on the Web. Once added to your Zotero database, the collection can be viewed in an iTunes-like interface, annotated, and searched. Items from Zotero can be added to documents by dragging them from Zotero into your document. Citations and bibliographies are created automatically using any one of hundreds of citation styles. Zotero syncs across devices, so changes made on your mobile device will update the Zotero file on your PC. You can share Zotero files with a work group, facilitating collaboration and resource sharing.

Mendeley (<http://www.mendeley.com/>). Mendeley has many features in common with Zotero. As with Zotero, in Mendeley personal reference databases are created from materials downloaded from the Web. These references can be searched and used to create citations and bibliographies in a variety of styles. The most innovative features of Mendeley relate to its use as a social network tool. While both Zotero and Mendeley allow the user to share personal libraries with a work group, Mendeley accommodates large private groups and public groups. Public groups allow you to join international communities of people following specific areas of interest. Mendeley will also recommend new references based upon your current Mendeley collection. By bringing a social network approach to bibliographic management, Mendeley provides a powerful tool for work groups and for consulting engineers interested in maintaining ties to their academic community.

7 FUTURE OF ONLINE ACCESS

Predicting the future is a tricky business and one rich with opportunity for embarrassment. That said, I have not done badly in comments made in previous editions of this handbook. As predicted, the OpenURL standard has become widely adopted in information systems, allowing an easy, intuitive connection between citation databases and digital copies of journal articles and conference papers. In the last edition of this book, I predicted that this feature would become commonplace within the next few years and now it is almost universal. I was also on target in suggesting that federated searching would continue to grow in popularity as a response to the ever-growing amount of information available online. While it provides some advantages, federated search has some inherent limitations and I suspect will never be more than a niche technology. My comments about the growing importance of institutional repositories also were fairly accurate. This area of development continues to grow and evolve and is addressed in more length below. While my analysis of industry trends proved pretty accurate, I did less well in

predicting a revival in interest in information agents. While I do believe that this approach deserves more investigation, the cost of doing so has inhibited progress in a period of global fiscal concern.

In looking at the coming decade, some trends seem well established. The dominance of the digital journal is now certain. Paper journals have become a novelty in most large academic libraries as online use has all but eclipsed use of the paper copy. While the move to eBooks has been slower, there is a clear trend toward greater eBook use. This is a factor of the greater trend toward publication of digital books as well as the result of the growing list of eBook readers and mobile devices with eReader apps. Expect to see this trend continue; however, while we may see the end of the paper publication of journals, I expect paper books to continue to be with us for some time to come.

On the search front, the future seems less clear. The current system of numerous databases of the technical literature does not serve the community well. There are so many options it is difficult to know where to search, even for information professionals. Federated searching, which amalgamates results from a number of databases into a single set of results, is plagued by duplicated entries, slow processing time, and inexact results. It will continue to be of some interest until a better solution can be found. Experiments are now underway at a number of universities using a Semantic Web to organize information. Under this approach articles would be indexed in a system that provided relational information, rather than a set of assigned keywords. Advocates of the approach see it as a more exact, more scalable system for cataloging a growing scholarly literature. The success of this model depends upon an automated approach for generating the relational indexing necessary for a Semantic Web approach. For a detailed review of the Semantic Web approach, interested readers can refer to Refs. 4 or 5 for detailed summaries of the approach.

In the last edition of this handbook I referred to the growing trend of universities to create digital repositories for their growing collections of digital materials. DSpace, an early software system for managing digital repositories, has been supplanted at a number of large universities by the Fedora-based Hydra and Islandora repository platforms. These repositories have benefited from a trend for universities to take over the digital publication of more of their intellectual property, rather than to hand it over to commercial publishers. So far this has most affected dissertation, theses, and technical reports. However, there are two business drivers in academia that may dramatically expand the use of the digital repository: MOOCs and data management requirements.

Recently universities began to produce massive open online courses (MOOCs). There is currently a great deal of excitement over MOOCs as a new approach to create massive online learning environments. Digital repositories are well positioned to benefit from the MOOCs and the online information they will create. While we are in the very early days of the MOOC, and I doubt that they will replace the traditional degree programs offered by universities, it seems clear that some aspects of university teaching will migrate to the online environment.

The other factor that will benefit from the presence of the digital repository is the growing list of requirements by grantors for researchers to make their data available to the public. The National Science Foundation (NSF) and the National Institutes of Health (NIH) now have imposed data-sharing requirements for their grants. A score of major governmental agencies and nongovernmental foundations have followed suit. When applying for a grant, investigators are required to file a data management plan (DMP) detailing how data will be preserved and made accessible to the commercial and academic research communities. The mission of long-term preservation and access fits perfectly with the goals of the digital repositories and the academic community has responded, led by the University of California, by the development of a DMP Tool (<https://dmp.cdlib.org/>). The DMP Tool assists investigators in preparing a data management plan and, often, directs them to a host university's digital repository as the site

to deposit research data. If university-based digital repositories become the default option for preserving data generated by research grants, this will become a major resource for industry and academia alike.

Beyond the growth of the digital repository, I anticipate that we will soon see efforts to link digital repositories into national, regional, and global networks. Existing implementation of the Shibboleth protocol (<http://shibboleth.net/>) have created secure collaborative environments across organizations. Expanded use of Shibboleth and interlinkages between digital repositories will create rich information environments for instruction and research. The landmark “Atkins” report⁶ on cyberinfrastructure provides an intriguing vision of the changes required in networks, computing, and information infrastructure to support science and engineering in the coming decades.

Given these trends, we can anticipate networks and technologies of increasing complexity which facilitate information products and services that are easier to use and more tailored to individuals’ needs and preferences. As more of the engineering literature moves to digital formats and publishers develop better economic models for selling online access, you can anticipate more availability of direct online access to engineering books and articles

8 OPTIONS FOR USING ONLINE INFORMATION

In this chapter I have outlined a variety of strategies for accessing online information and reviewed a number of content and service providers and a sampling of databases. By using these tools you can create your own unique desktop library, with access to the global literatures of engineering, business and trade news, intellectual property, government R&D, and much more. Never in the history of humankind has so much information been so readily available to so many.

What does this mean for you as an engineer? It creates an unparalleled opportunity to:

- Investigate unfamiliar technologies for new projects
- Determine the best technologies and suppliers to use for an assignment
- Research the latest trends in a technical area
- Get background material on a person, company, or event prior to a meeting
- Find the latest version of an industry standard or military specification
- Locate patents for background on proprietary technologies
- Research a topic—for a quick update or a comprehensive review

The bottom line is that online information research saves you time, gives you more technical options, helps you avoid “reinventing the wheel,” and gives you an edge on your competition by making you better informed. This is true whether you are preparing a grant proposal or involved in a “fire-fighting” emergency in an industrial setting. If you are missing information or using slow, traditional means of finding it, you are putting yourself and your organization at a competitive disadvantage!

In this chapter I have provided a summary of some of the information resources available to you online and outlined several strategies for how you can effectively use them. As information resources continue to grow and evolve, fast easy access to information is only a fraction of the advantage you obtain over more traditional approaches. By putting the necessary information into your hands— anytime and anywhere you need to be—online information can get you started on a new project or help you survive an ISO 9000 audit. By linking you to online communities, online access helps connect you to your profession in a way unimaginable even a few years ago.

REFERENCES

1. T. Allen, *Managing the Flow of Technology: Technical Transfer and the Dissemination of Technical Information Within the R&D Organization*, MIT Press, Cambridge, MA, 1977.
2. M. Holland, "Engineering," in N. Pruett (Ed.), *Scientific and Technical Libraries*, Vol. 1: *Functions and Management*, Academic, New York, 1986, pp. 119–142.
3. T. Pinelli, "Distinguishing Engineers from Scientists—The Case for an Engineering Knowledge Community," *Sci. Technol. Libraries*, **21**(3/4), 131–163, 2001.
4. Z. Ahmed, "Review: Semantic Web; Ontology Specific Languages for Web Application Development," 2012, available: http://zeeshanahmed.info/pdf%20files/zeeshan_sw_vol4_iss2_ijwa.pdf.
5. B. Sean, "The Semantic Web: An Introduction," 2007, available: <http://infomesh.net/2001/swintro>.
6. D. Atkins (chair), "Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Advisory Panel on Cyberinfrastructure," February 3, 2003, available: <http://www.nsf.gov/od/oci/reports/atkins.pdf>.

CHAPTER 28

SOURCES OF MECHANICAL ENGINEERING INFORMATION

Fritz Dusold

Mid-Manhattan Library Science and Business Department (Retired)
New York, New York

Myer Kutz

Myer Kutz Associates, Inc.
Delmar, New York

1 INTRODUCTION	823	5 CODES, SPECIFICATIONS, AND STANDARDS	826
2 PRIMARY LITERATURE	823	6 GOVERNMENT PUBLICATIONS	827
2.1 Periodicals	824	7 ENGINEERING SOCIETIES	827
2.2 Conference Proceedings	824	8 LIBRARIES	828
3 INDEXES AND ABSTRACTS	824	REFERENCES	828
3.1 Manual Searching	824		
3.2 Online Searching	825		
4 ENCYCLOPEDIAS AND HANDBOOKS	825		

1 INTRODUCTION

This chapter is designed to enable the engineer to find information efficiently and to take advantage of all available information. The emphasis is placed on publications and services designed to identify and obtain information. Because of space limitations, references to individual works which contain the required information are limited to a few outstanding or unusual items.

2 PRIMARY LITERATURE

The most important source of information is the primary literature. It consists mainly of the articles published in periodicals and of papers presented at conferences. New discoveries are first reported in the primary literature. It is, therefore, a major source of current information. Peer review and editorial scrutiny, prior to publication of an article, are imposed to ensure that the article passes a standard of quality. Most engineers are familiar with a few publications but are not aware of the extent of the total production of primary literature. *Engineering Index*¹ (known as *Compendex* in its electronic version) alone abstracts material from thousands of periodicals and conferences.

Handbooks and encyclopedias are part of the secondary literature. They are derived from primary sources and make frequent references to periodicals. Handbooks and encyclopedias

are arranged to present related materials in an organized fashion and provide quick access to information in a condensed form.

While monographs—books written for professionals—are either primary or secondary sources of knowledge and information, textbooks are part of the tertiary literature. They are derived from primary and secondary sources. Textbooks provide extensive explanations and proofs for the material covered to provide the student with an opportunity to understand a subject thoroughly.

2.1 Periodicals

In most periodicals published by societies and commercial publishers, articles are identified usually by issue and/or volume, date, and page number. Bibliographic control is excellent, and it is usually a routing matter to obtain a copy of a desired article. But some problems do exist. The two most common are periodicals that are known by more than one name and the use of nonstandard abbreviations. Both of these problems could be solved by using the International Standard Serial Number (ISSN), which accurately identifies each publication. The increasing size and use of automated databases should provide an impetus to increased use of ISSN or some other standard.

The first scientific periodical, *Le Journal des Scavans*, was published January 4, 1665. The second, *Philosophical Transactions of the Royal Society (London)*, appeared on March 6, 1665. The number of scientific periodicals has been increasing steadily with some setbacks caused by wars and natural catastrophies. The accumulated body of knowledge is tremendous. Much of this information can be retrieved by consulting indexes, abstracts, and bibliographies.

2.2 Conference Proceedings

The bibliographic control for papers presented at conferences is not nearly as good as for periodicals. The responsibility for publishing the papers usually falls upon the sponsoring agent or host group. For major conferences the sponsoring agency is frequently a professional society or a department of a university. In these cases an individual with some experience in publishing is usually found to act as an editor of the proceedings. In other instances the papers are issued prior to the conference as preprints. In still other situations the papers will be published in a periodical as a special issue or distributed over several issues of one or more periodicals. An additional, unknown, percentage of papers are never published and are only available in manuscript form from the author.

3 INDEXES AND ABSTRACTS

Toward the end of the last century the periodical literature had reached a volume that made it impossible for the “educated man” to review all publications. In order to retrieve the desired or needed information, indexes and abstracts were prepared by individual libraries and professional societies. In the 1960s computers became available for storing and manipulating information. This led to the creation and marketing of automated data banks.

3.1 Manual Searching

The major abstracts typically provide the name of the author, a brief abstract of the article, and the title of the article and identify where the article was published. Alphabetical author and subject indexes are usually provided, and a unique number is assigned to refer to the abstract.

Many abstracts are published monthly or more frequently. Annual cumulations are available in many cases. The most important abstracts for engineers are:

*Compendex*¹

*Science Abstracts*²

Series A: Physics Abstracts

Series B: Electrical and Electronics Abstracts

Series V: Computer and Control Abstracts

*Chemical Abstracts*³

*Metals Abstracts*⁴

A comprehensive listing of abstracts and indexes can be found in *Ulrich's International Periodical Directory*.⁵

3.2 Online Searching

Most of the major indexes and abstracts are now available online. For a comprehensive list of databases and online vendors see *Information Industry Market Place*.⁶ In addition to indexes and abstracts, periodicals, encyclopedias, and handbooks are available online. There seems to be virtually no limit to the information that can be made available online. The high demand for quick information retrieval ensures the expansion of this service.

In addition to the online indexes, several library networks and consortia, such as OCLC, the Online Computer Library Center, located in Columbus, Ohio, produce online databases. These are essentially equivalent to the catalogs of member libraries and can be used to determine which library owns a particular book or subscribes to a particular periodical.

4 ENCYCLOPEDIAS AND HANDBOOKS

There are well over 300 encyclopedias and handbooks covering science and technology. "amazon.com" and "barnesandnoble.com" are Internet sites with comprehensive catalogs of books. The date of publication should be checked before using any of these works if the required information is likely to have been affected by recent progress. The following list represents only a sampling of available works of outstanding value.

*Kirk-Othmer Encyclopedia of Chemical Technology*⁷ provides a comprehensive and authoritative treatment of a wide range of subjects, with heaviest concentration on materials and processes. The basic set is updated by supplements.

*Encyclopedia of Polymer Science and Engineering*⁸ is one of the major works in this important area of materials.

*Metals Handbook*⁹ provides an encyclopedic treatment of metallurgy and related subjects. Each of the volumes is devoted to a separate topic such as mechanical testing, powder metallurgy, and heat treating. Each of the articles is written by a committee of experts on that particular topic.

CRC Handbook of Chemistry and Physics,¹⁰ popularly known as the "Rubber Handbook," is probably the most widely available handbook. It is updated annually to include new materials and to provide more accurate information on previously published sections as soon as the information becomes available.

The increasing concern with industrial health and safety has placed an additional responsibility on the engineer to see that materials are handled in a safe manner. Sax's *Dangerous*

*Properties of Industrial Materials*¹² provides an authoritative treatment of this subject. This book also covers handling and shipping regulations for a large variety of materials.

Engineers have always been concerned with interaction between humans and machines. This area has become increasingly sophisticated and specialized. *Human Factors and Ergonomics Design Handbook*¹² is written for the design engineer rather than the human factor specialist. The book provides the engineer with guidelines for designing products for convenient use by people. Another important title is *Handbook of Human Factors and Ergonomics*.¹³

Engineering work frequently requires a variety of calculations. *Standard Handbook of Engineering Calculations*¹⁴ provides the answer to most problems. Although most of the information in this handbook is easily adaptable to computer programming, a new edition would probably take greater advantage of the increasing availability of computers.

Conservation of energy remains an important consideration, owing to energy's increasing cost. Two titles dealing with this subject are *Handbook of Energy Efficiency and Renewable Energy*¹⁵ and *Energy Management Handbook*.¹⁶

Composite materials frequently offer advantages in properties and economy over conventional materials. Information about composites can be found in several handbooks.

When England converted to the metric system, the British Standards Institution published *Metric Standards for Engineers*.¹⁷ With the increasing worldwide distribution of products, metric units will gain in importance regardless of the official position of the U.S. government. This handbook offers the engineer an authoritative and detailed treatment of metrification.

5 CODES, SPECIFICATIONS, AND STANDARDS

Codes, specifications, and standards are produced by government agencies, professional societies, businesses, and organizations devoted almost exclusively to the production of standards. In the United States the American National Standards Institute (ANSI, New York) acts as a clearinghouse for industrial standards. ANSI frequently represents the interests of U.S. industries at international meetings. Copies of standards from most industrial countries can be purchased from ANSI as well as from the issuer.

Copies of standards issued by government agencies are usually supplied by the agency along with the contract. They are also available from several centers maintained by the government for the distribution of publications. Most libraries do not collect government specifications.

Many of the major engineering societies issue specifications in areas related to their functions. These specifications are usually developed and revised by membership committees.

The American Society of Mechanical Engineers (ASME, New York) has been a pioneer in publishing codes concerned with areas in which mechanical engineers are active. In 1885 ASME formed a Standardization Committee on Pipe and Pipe Threads to provide for greater interchangeability. In 1911 the Boiler Code Committee was formed to enhance the safety of boiler operation. The 1983 ASME Boiler and Pressure Vessel Code was published in a metric (SI) edition, in addition to the edition using U.S. customary units, to reflect its increasing worldwide acceptance. The boiler code covers the design, materials, manufacture, installation, operation, and inspection of boilers and pressure vessels. Revisions, additions, and deletions to the code are published twice yearly during the three-year cycle of the code.

A frequently used collection of specifications is *Annual Book of Standards*¹⁸ issued by the American Society for Testing and Materials (ASTM). These standards are prepared by committees drawn primarily from the industry most immediately concerned with the topic.

The standards written by individual companies are usually prepared by a member of the standards department. They are frequently almost identical to standards issued by societies and

government agencies and make frequent references to these standards. The main reason for these “in-house” standards is to enable the company to revise a standard quickly in order to impose special requirements on a vendor.

The large number of standards issued by a variety of organizations has resulted in a number of identical or equivalent standards. Information Handling Services (IHS, Englewood, CO) makes available virtually all standards on CD-ROM.

6 GOVERNMENT PUBLICATIONS

The U.S. government is probably the largest publisher in the world. Most of the publications are available from the Superintendent of Documents (U.S. Government Printing Office, Washington, D.C.). Publication catalogs are available on the Government Printing Office website, GPO.gov. Increasingly, the GPO is relying on electronic dissemination rather than print. These publications are provided, free of charge, to depository libraries throughout the country. Depository libraries are obligated to keep these publications for a minimum of five years and to make them readily available to the public.

The government agencies most likely to publish information of interest to engineers are probably the National Institute of Science and Technology, the Geological Survey, the National Oceanic and Atmospheric Administration, and the National Technical Information Service.

7 ENGINEERING SOCIETIES

Engineering societies have exerted a strong influence on the development of the profession. The ASME publishes the following periodicals in order to keep individuals informed of new developments and to communicate other important information:

Applied Mechanics Reviews (monthly)

CIME (*Computers in Mechanical Engineering*, published by Springer-Verlag, New York)

Mechanical Engineering (monthly)

Transactions (quarterly)

The transactions cover the following fields: power, turbomachinery, industry, heat transfer, applied mechanics, bioengineering, energy resources technology, solar energy engineering, dynamic systems, measurement and control, fluids engineering, engineering materials and technology, pressure vessel technology, and tribology.

Many engineering societies have prepared a code of ethics in order to guide and protect engineers. Societies frequently represent the interests of the profession at government hearings and keep the public informed on important issues. They also provide an opportunity for continuing education, particularly for preparing for professional engineers examinations. The major societies and trade associations in the United States are:

American Concrete Institute

American Institute of Chemical Engineers

American Institute of Steel Construction

American Society of Civil Engineers

American Society of Heating, Refrigerating, and Air-Conditioning Engineers

American Society of Mechanical Engineers

Institute of Electrical and Electronics Engineers

Instrument Society of America
National Association of Corrosion Engineers
National Electrical Manufacturers Association
National Fire Protection Association
Society of Automotive Engineers
Technical Association for the Pulp and Paper Industry
Underwriters Laboratories

8 LIBRARIES

The most comprehensive collections of engineering information can be found at large research libraries. Four of the largest in the United States are

John Crerar Library
35 West 33rd Street
Chicago, IL 60616
Library of Congress
Washington, DC 20540

Linda Hall Library
5109 Cherry Street
Kansas City, MO 64110

New York Public Library—Science, Industry and Business Library
188 Madison Avenue
New York, NY 10016

These libraries are accessible to the public. They do provide duplicating services and will answer telephoned or written reference questions.

Substantial collections also exist at universities and engineering schools. These libraries are intended for use by faculty and students, but outsiders can frequently obtain permission to use these libraries by appointment, upon payment of a library fee, or through a cooperative arrangement with a public library.

Special libraries in business and industry frequently have excellent collections on the subjects most directly related to their activity. They are usually only available for use by employees and the company.

Public libraries vary considerably in size, and the collection will usually reflect the special interests of the community. Central libraries, particularly in large cities, may have a considerable collection of engineering books and periodicals. Online searching is becoming an increasingly frequent service provided by public libraries.

Regardless of the size of a library, the reference librarian should prove helpful in obtaining materials not locally available. These services include interlibrary loans from networks, issuing of courtesy cards to provide access to nonpublic libraries, and providing the location of the nearest library that owns needed materials.

REFERENCES

1. *Compendex*, Elsevier.
2. *Science Abstracts*, INSPEC, Institution of Electrical Engineers.

3. *Chemical Abstracts*, American Chemical Society.
4. *Metals Abstracts*, Metals Information.
5. *Ulrich's International Periodicals Directory*, R. R. Bowker, New York.
6. *Information Industry Market Place*, An International Directory of Information Products and Services, R. R. Bowker, New York.
7. *Kirk–Othmer Encyclopedia of Chemical Technology*, 5th ed., Wiley-Interscience, Hoboken, NJ, 2007 (26 vol6).
8. *Encyclopedia of Polymer Science and Engineering*, 2nd ed., Wiley-Interscience, Hoboken, NJ 2014 (15 vols.).
9. *Metals Handbook*, American Society for Metals, Metals Park, OH.
10. *CRC Handbook of Chemistry and Physics*, CRC Press, Boca Raton, FL (annual).
11. Sax's *Dangerous Properties of Industrial Materials*, Wiley, Hoboken, NJ, 2012 (5 vols).
12. B. Tillman et al, *Human Factors and Ergonomics Design Handbook*, McGraw-Hill, New York, 2015
13. G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics*, Wiley, Hoboken, NJ, 2012.
14. T. G. Hicks et al., *Standard Handbook of Engineering Calculations*, McGraw-Hill, New York, 2014.
15. F. Kreith and D. Y. Goswami, *Handbook of Energy Efficiency and Renewable Energy*, CRC Press, Boca Raton, FL, 2007.
16. W. C. Turner, *Energy Management Handbook*, Fairmont Press, Lisburn, GA, 2012.
17. *Metric Standards for Engineers*, British Standards Institution, London, 1967.
18. *Annual Book of Standards*, American Society for Testing and Materials, West Conshohocken, PA.

Index

A

- ABC (activity-based costing), 23–24
- Abrasive barrel finishing, 229
- Abrasive belt finishing, 230
- Abrasive disks, 152–153
- Abrasive flow machining (AFM), 158–159, 162
- Abrasive jet machining (AJM), 159, 162
- Abrasive machines, 157
- Abrasive machining, 153–158
 - abrasives, 154–156
 - temperature, 156–158
- Abstracts, 824–825
- AC (adaptive control), 347, 348
- Acceptance sampling, 335–336
- Accident prevention, 727–729
- Accreditation (quality management systems), 668–669
- Active agents, 550–552
- Activity-based costing (ABC), 23–24
- Activity network diagram (AND), 631, 634, 650, 651
- Adaptability, 14
- Adaptive control (AC), 347, 348
- Adhesive bond testing:
 - bond testers, 464–465
 - NDE case study, 474–476
- ADLI (approach, deployment, learning, and integration) model, 681–682
- ADR (alternative dispute resolution), for patent disputes, 800
- Adware, 806
- AED (automated external defibrillator) training, 735
- Aerospace:
 - braided carbon rope seals, 294–295
 - noncontacting seals, 304–308
 - space organizations:
 - AS 9100 quality standard, 670, 672, 673
 - mission assurance, 674–675
- AeroSTEP, 392
- Affinity diagram, 629, 630
- AFM (abrasive flow machining), 158–159, 162
- After-market product hazards, 768–770
- After tax analysis, 601
- Agency (employee liability), 752–753
- Agents, 520–531
 - active, sleeping, and locked, 550–552
 - architectures, 523–526
 - coalitions, 529–530
 - cognitive, 523
 - defined, 521
 - deliberative, 523, 524
 - federations, 530
 - floating, 552–554
 - hierarchies, 529
 - holarchies, 529
 - holonic, 518
 - intelligent, 522
 - interagent communication, 526–528
 - markets, 530
 - matrix, 531
 - multiagent systems, 520–523
 - organization of, 528–531
 - predecessor validation, 552, 553
 - reactive, 523
 - societies, 530
 - teams, 530
 - types of, 523
- Aggregate planning (AP), 73–77
 - approaches to, 74–75
 - costs of, 74
 - lack of adoption of, 76–77
 - levels of aggregation/disaggregation, 76
 - meeting demand fluctuations, 74
- Aggregation levels (production planning), 76
- Agility, 519
- AGVs (automated guided vehicles), 505
- AIAG (Automotive Industry Advisory Group), 673
- Aircraft turbine engines:
 - brush seals, 317, 319
 - packings and braided rope seals, 293
- Air pollutants, 301. *See also* Emissions control
- AJB (abrasive jet machining), 159, 162
- Alliance for Performance Excellence, 684
- Allied Signal, 392
- Allison Engines, 319
- Alternative dispute resolution (ADR), for patent disputes, 800
- Altshuller, Genrich, 362–364, 370–371, 374
- Aluminum oxide (abrasive), 154

- American Health Care Association Quality Award, 686
- American National Standards Institute (ANSI), 826
- American Petroleum Industry (API), 302
- American Society for Quality (ASQ), 680, 684
- American Society for Testing and Materials (ASTM), 826
- American Society of Mechanical Engineers (ASME), 818, 826, 827
- ANAB (ANSI-ASQ National Accreditation Board), 670, 675
- Analyze phase (DMADVV), 660–661
- Analyze phase (DMAIIC), 643–649
- AND, *see* Activity network diagram
- Angle bending (metals), 212
- Annual Book of Standards*, 826
- Annual worth (AW), 588, 592–595
- Annular seals, 305, 308, 313
- Anodization system material handling
 - (case study), 543–550
 - agent design, 546–548
 - interactions, 548–550
 - PACO approach, 545–546
- Anodizing, 232–233
- ANSI (American National Standards Institute), 826
- ANSI-ASQ National Accreditation Board (ANAB), 670, 675
- AOQ (average outgoing quality level), 335
- AP, *see* Aggregate planning
- API (American Petroleum Industry), 302
- Approach, deployment, learning, and integration (ADLI) model, 681–682
- Arc evaporation, 241–242
- Arc-welding processes, 41–43
- Argon dc glow discharge, 236–237
- ARIZ (inventive problem solving), 383–388
 - caveat for, 388
 - steps in, 384–385
- Artificial intelligence, 520–521. *See also* Intelligent control of MHSs
- Artisan guilds, 5
- Art of engineering, 749–751
- ASME, *see* American Society of Mechanical Engineers
- ASME Boiler and Pressure Vessel Code, 826
- ASQ (American Society for Quality), 680, 684
- AS quality standards:
 - AS9100: aviation, space, and defense organizations, 670, 672–675
 - AS9102: Aerospace First Article Inspection Requirement, 672
 - AS9103: Variation Management of Key Characteristics, 672
 - AS9110: Quality Management Systems—Requirements for Aviation Maintenance Organizations, 672
 - AS9120: Quality Management Systems—Requirements for Aviation, Space and Defense Distributors, 672
 - AS9131: Quality Systems Non-Conformance Documentation, 672
- AS/RSs (automated storage and retrieval systems), 505
- Assembly line balancing, 97–104
 - definitions related to, 97–98
 - design of assembly line, 100
 - mixed-model assembly lines, 101–103
 - parallel line balancing, 103–104
 - structure of problem, 99
 - techniques for, 100–101
- Assembly line design, 100
- Assessment of manufacturing systems, 189–193
 - data collection, 190–191
 - planning, 189–190
 - reporting and project formulation, 192–193
 - site visit and inspection, 191
- Assumption of risk, 767–768
- ASTM (American Society for Testing and Materials), 826
- Attitude, success as function of, 567
- AT&T Power Systems, 680
- Attribute characteristics, 400
- AT&T Western Electric, 655
- Authority:
 - as basis of interpersonal power, 569
 - and employee liability, 752–753
- Automated external defibrillator (AED) training, 735
- Automated guided vehicles (AGVs), 505
- Automated storage and retrieval systems (AS/RSs), 505
- Automation, 340–341, 516. *See also* Intelligent control of MHSs
- Automation principle (material handling), 502–503
- Automatization, 340
- Automotive industry, 516
 - alternative fuel vehicles, 582
 - ISO 9001 version for, 672, 673
 - ISO/TS 16949 quality standard, 672, 673
- Automotive Industry Advisory Group (AIAG), 673
- Autonomous vehicle storage and retrieval systems (AVS/RSs), 505
 - case study, 510–511
 - material handling, 505
- Available units, 77, 81
- Average outgoing quality level (AOQ), 335

- Aviation, 293
 - AS 9100 quality standard for organizations, 670, 672, 673
 - aircraft turbine engines, 293, 317, 319
- AVS/RSSs, *see* Autonomous vehicle storage and retrieval systems
- AW (annual worth), 588, 592–595

- B**
- Back-order, 66
- Back pitch (joints), 262
- Baggage handling system (case study), 531–544
 - agent design, 534
 - agent interactions and ontology, 538–540
 - internal agent reasoning, 540–543
 - performance criteria, 532–533
 - toploader, 534–538
 - worst-case scenario, 533–534
- Balance delay of a workstation, 97, 99
- Baldrige, Malcolm, 680
- Baldrige Awards, *see* Malcolm Baldrige Awards
- Baldrige National Quality Award (BNQA), 668, 670, 680–684, 686, 687
- Band saws, 152–153
- Barrel finishing, 229
- Batch production, 342
- BCR (benefit–cost ratio), 597–598
- BDI agent architecture, 524
- Beam processes (PVD), 247–249
- Bearing failure (joints), 264, 265
- Bearing-type connections (joints), 262–268
- Benchmarking, 15, 566
- Bend allowances (metals), 211
- Bending (metals), 208, 210–213
- Bending force (metals), 212
- Bending reduction, design of mechanical fasteners for, 281
- Benefit–cost ratio (BCR), 597–598
- Ben & Jerry's, 18
- Best mode requirement (patents), 786
- Bhutan, 25
- Biomimicry, 18–19
- Bipolar pulsed power (sputtering), 244
- Blanking, 213, 214
- Blasting, 229
- Blow molding (plastics), 225–226
- BNQA, *see* Baldrige National Quality Award
- Boeing, 20, 256, 392
- Boid model, 524
- Bolts, 261–262. *See also* Mechanical fasteners stresses, 271–272
 - structural purpose of, 269
 - tightening:
 - torque and turn together method, 278–279
 - turn-of-nut, 277–278
 - ultrasonic measurement of bolt stretch or tension, 279–280
- Bolted and riveted joints:
 - butt joints, 262–264, 266–268
 - clamping force upper limits, 271–272
 - efficiency, 264
 - friction-type connections, 268–270
 - lap joints, 262, 264–265
 - slip characteristics evaluation, 276–277
 - theoretical behavior of joint under tensile loads, 272–276
 - types, 262–264
- Bolting paradigm, 256
- Bolt load (gaskets), 287
- Bolt pattern (gaskets), 288–289
- Bonderizing, 233
- Bonding materials, grinding and, 154
- Bond testing, 464–465
- Boring, 129
- Brainstorming, 15, 650
- Brazing, 41, 45
- Breach of contract, 756
- Break-even analysis, 602
- Break-even conditions (metal cutting), 132
- Bristle stiffening (brush seals), 316
- British Library Document Supply Services, 818
- Broaching, 122, 148–151
- Brush seals, 313–319
 - aircraft turbine engines, 317, 319
 - bristle stiffening, 316
 - brush pack considerations, 315
 - cryogenic, 317
 - design considerations, 314–315
 - flow modeling, 317–318
 - ground-based turbine engines, 319
 - hybrid, 319
 - implementation issues, 315
 - leakage performance, 316–317
 - materials for, 318–319
 - pressure closing, 316
 - seal hysteresis, 315, 316
- Buffing, 158, 230
- Building information management, 395
- Burnishing, 210
- Business liability, 755–757
 - contractual obligations, 755–756
 - insurance for individual engineers, 756
 - negligence for services, 755
- Business process management, 562–563
- Butt joints, 262–264
- Butt welding, 207

C

- CAA (Clean Air Act, 1990), 302, 696
- CAD, *see* Computer-aided design
- CAD/CAM, 346–347, 351
- CAI (computer-aided inspection), 340
- CAM, *see* Computer-aided manufacturing
- CAM (computer-aided machining), 346–347
- Cameras, infrared, 468–469
- Canada, filing patents in, 801
- Canadian runner molding, 47
- Capability Maturity Model Integration (CMMI®), 668, 675, 676
- Capital expenditure decisions, 583
- Capitalized cost (CC), 599
- Capital recovery (CR), 598
- CAPP (computer-aided process planning), 358
- CAPP (computer-aided production planning), 340
- Carbides, 126–127, 131
- Carnegie Mellon University (CMU), 675
- Cartesian coordinate system, 345–346
- Cash flows, 583–584, 588–591
- Cash flow diagrams, 584–587
- Casting:
 - metals, *see* Metal casting and molding
 - threads, 146
- Cast nonferrous alloys, 126
- Catastrophe analysis, 701
- Causal forecasting methods, 58–60
- Cause and effect, traditional approach to, 400
- Cause-and-effect diagram, 625, 643–644
- Cause of action (lawsuits), 751
- CBN (cubic boron nitride), 127, 154
- CC (capitalized cost), 599
- c* charts, 334–335
- Centrifugal casting (metals), 220–221
- Ceramic process, 38, 224
- Ceramic tool inserts, 127
- CERCLA (Comprehensive Environmental Response, Compensation, and Liability Act, 1980), 695–696
- Certification (quality management systems), 668–670
- Challenges, motivation as function of, 566–567
- Change management, 653
- Channel bending, 212
- Charter, project, 637–638, 656
- Check sheets, 625, 626
- Chemical blanking, 178
- Chemical conversions (surface treatment), 232–233
- Chemical machining (CHM), 178
- Chemical milling, 178
- Chemical oxide coatings, 233
- Chemical vapor deposition (CVD), 235
- Chesapeake Paper Products, 757
- CHM (chemical machining), 178
- Chromate coatings, 233
- CIMS, *see* Computer-integrated manufacturing system
- Circular saws, 152–153
- Classification of hazards, 700
- Clausing, Donald, 657
- Clean Air Act (CAA, 1990), 302, 696
- Clean Air Act Amendment, 301
- Cleaning (surface treatment), 35, 227–230
- Clean Water Act (CWA, 1970), 694
- Clean Water State Revolving Fund, 694
- Client (software), 806
- Clothing, safety, *see* Personal protective equipment (PPE)
- Cloud computing, 14, 20, 806
- CMMI®, *see* Capability Maturity Model Integration
- CMU (Carnegie Mellon University), 675
- Coalitions (agents), 529–530
- Coated tools, 127
- Coatings, 230–232
 - chemical conversions, 233
 - physical vapor deposition, 235–249
 - chemical vapor deposition vs., 235
 - film formation and growth, 237, 239
 - glow discharge plasma, 236–238
 - processes, 239–249
 - temporary corrosion protection, 232
- Codes, 826
- Cognitive agents, 523, 524
- Coining, 210
- Coinjection molding (plastics), 225
- COI (cube per order index) storage policy, 501, 502
- Cold-chamber die casting, 37, 223
- Cold drawing (metals), 215–216
- Cold forging (metals), 209
- Cold roll forming (metals), 212
- Cold rolling (metals), 209
- Cold spinning (metals), 216
- Cold-working processes (metals), 207–217
 - bending, 208, 210–213
 - classification of, 208
 - drawing, 208, 215–217
 - shearing, 208, 213–215
 - squeezing, 208–210
- Collective multifunctional evaluations, 612
- Combined loads (bolts), 271
- Combustion machining, 179
- Commitment, building, 576
- Common-cause variation, 654
- The Commons, 18

- Communication channels, 574
- Comparative negligence, 766, 768
- COMPENDEX, 815, 823
- Complex butt joint (bearing-type connection), 266–268
- Compliant features (fastening), 258–259
- Compositions of matter (legal term), 778
- Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA, 1980), 695–696
- Computed tomography (CT), 455–456
- Computer-aided assembly, 516, 517
- Computer-aided design (CAD):
 - in CIMS, 339–340
 - NC manufacturing systems, 346–347
 - rapid prototyping, 180
 - STEP model, 392, 395
- Computer-aided inspection (CAI), 340
- Computer-aided machining (CAM), 47
- Computer-aided manufacturing (CAM), 346–347
 - in CIMS, 339–340
 - STEP-NC, 392, 394, 395
- Computer-aided process planning (CAPP), 358
- Computer-aided production planning (CAPP), 340
- Computer analysis tools:
 - annular seals, 305, 308
 - labyrinth seals, 311
 - mechanical face seals, 305, 308
- Computer-integrated manufacturing system (CIMS), 339–358
 - computers in, 351–353
 - CNC and DNC systems, 352
 - flexible manufacturing systems, 352–353
 - hierarchical computer control, 351–352
 - manufacturing cells, 352
 - deficits of, 517
 - definitions and classifications, 340–344
 - automation, 340–341
 - models for production operations, 342–344
 - production operations, 341–342
 - production plants, 342
 - group technology, 353–358
 - computer-aided process planning, 358
 - machine cell designs, 357–358
 - part family formation, 354
 - parts classification and coding, 354–357
 - production flow analysis, 357
 - industrial robots, 349–351
 - applications, 350–351
 - configurations, 349, 350
 - control and programming, 349, 350
 - defined, 349
 - limitations of, 520
 - numerical control manufacturing systems, 344–348
 - adaptive control, 347, 348
 - CAD/CAM part programming, 346–347
 - coordinate system, 345–346
 - machinability data prediction, 347
 - numerical control components, 344–345
 - parts selection for machining, 346
 - programming by scanning and digitizing, 347
 - Concept generation (DMADV), 661
 - Conditioned water (cleaning), 229
 - Conference proceedings, 824
 - Conflict management, 576
 - Conservation, 19
 - Constant demand inventory models, 65
 - Continuation-in-part patent applications, 794–795
 - Continuation patent applications, 794–795
 - Continuing duty to warn, 768–770
 - Continuous improvement, 4, 10
 - Continuous-path (contouring) NC/CNC systems, 352
 - coordinate system, 345–346
 - hierarchical computer control, 351
 - Continuous-path robots, 349, 350
 - Contouring robots, 349
 - Contractual obligations:
 - business liability, 755–756
 - employment agreements, 753–754
 - liability under law of agency, 753
 - Contradictions (in TRIZ), 388
 - contradiction matrix, 371–373
 - contradictions principle, 365–367
 - problems without, 376–378
 - Contributory infringement (patents), 797
 - Contributory negligence, 766, 768
 - Control, intelligent, *see* Intelligent control of MHSs
 - Control charts, 326–331, 625, 626
 - for attributes, 332–336
 - c* charts, 334–335
 - DMAIIC control phase, 653–656
 - np* charts, 333–334
 - operation characteristic curves, 332
 - p* charts, 332–334
 - u* charts, 334–335
 - \bar{X} , *R*, and σ , 326–331
 - Control phase (DMAIIC), 653–656
 - Control plan, 653, 663
 - Control systems level (manufacturing systems), 186, 188
 - Control variables, 398, 399, 401–404
 - Converters, 341
 - Conveyors, 504
 - Cookies, 806

- Cooperation, 519
 - Coordinate system (NC systems), 345–346
 - Coordination (agent communication), 526–527
 - Core competencies, 20
 - Correlations:
 - TRA, 412, 414, 415
 - traditional approach to, 401
 - Corrosion, 261, 281
 - Corrosion protection, 230–233
 - Cost(s):
 - aggregate planning, 74
 - of energy, 184, 185
 - in forecasting, 63
 - inventory, 63–65
 - life-cycle, 503
 - Cost analysis, 23–24
 - Cost–benefit (CB), 610
 - Cost/benefit analysis, 649–650
 - Cost reduction decisions, 583
 - CR (capital recovery), 598
 - Cranes, 505
 - CRC Handbook of Chemistry and Physics*
 (“Rubber Handbook”), 825
 - Creativity, *see* TRIZ
 - Crevice corrosion, 261
 - Critical to quality analysis (CTQ cascade), 650, 652
 - Cryogenic brush seals, 317
 - CT (computed tomography), 455–456
 - CTQ cascade (critical to quality analysis), 650, 652
 - Cube per order index (COI) storage policy, 501, 502
 - Cubic boron nitride (CBN), 127, 154
 - Cutoff (shearing), 215
 - Cutting:
 - gear form cutting, 144–146
 - metal-cutting processes:
 - economics of, 123, 125–126
 - environmentally benign, 31
 - principles, 116–119
 - plastics, 153
 - threads, 146–148
 - Cutting fluids, 31–32, 128
 - Cutting forces, 119–121
 - measurement of, 119
 - square-end punches and dies, 214
 - Cutting-off processes, 152–153
 - Cutting speeds, 129
 - internal threads, 146, 147
 - by material, 131
 - for maximum production, 126
 - milling, 142
 - for minimum cost, 126
 - planing and shaping, 149
 - and tool life, 124
 - Cutting-tool materials, 126–129
 - CVD (chemical vapor deposition), 235
 - CWA (Clean Water Act, 1970), 694
 - Cycle time, 97
 - Cyclical movements (time series), 60
 - Cyclical tension loads (mechanical fasteners), 280–282
- ## D
- DAI (distributed artificial intelligence), 521
 - Dangerous Properties of Industrial Materials*
 (Sax), 825–826
 - Database, 806
 - Database services, 813–817
 - fee-based, 815–816
 - free, 816–817
 - for mechanical engineers, 814–815
 - Data collection (systems assessment), 189–191
 - Data collection forms, 639
 - Data collection plan, 639
 - Datafile, 807
 - Data fusion, 443
 - Data management plan, 807
 - dc diode sputtering (PVD), 242–243
 - DDB (double declining balance) depreciation, 600
 - Deburring:
 - electrochemical, 166
 - thermal, 179
 - Decision analysis, 701
 - Decision making, 187, 606, 611
 - Declaratory judgment actions (patents), 799
 - Deep drawing (metals), 216
 - Defects, *see* Product defects
 - Defendant (lawsuits), 751
 - Defenses (legal):
 - to patent enforcement, 798
 - to product liability, 764–768
 - assumption of risk, 767–768
 - contributory/comparative negligence, 766
 - safety standards’ role in, 765–766
 - state of the art, 764–765
 - Defense organizations:
 - AS 9100 quality standard, 670, 672, 673
 - mission assurance, 674–675
 - Define phase (DMADV), 656–657
 - Define phase (DMAIIC), 637–639
 - Deliberative agents, 523, 524
 - Dell, 20–21
 - Dell, Michael, 21
 - Delphi method, 55

- Demand:
 - defined, 65
 - inventory, 64
 - meeting fluctuations in, 74
 - planning to meet fluctuations in, 74
- Demand during lead time, 66
- Deming, W. Edwards, 12, 15, 621, 678
- Deming Prize, 678–680, 686, 687
- Dependencies (agents), 526–527
- Depreciation, 599–600
- Description requirement (patents), 786
- Descriptor (databases), 807
- Deseasonalization, 54, 61, 62
- Design, 15
 - documentation of process, 770–771
 - safety engineering, 719–724
 - for sustainability, 18
 - technical and financial considerations, 591
- Design application (patents), 775, 776
- Design Excellence, 636. *See also* Total quality management (TQM)
- Design flaws, 760–761
- Design hierarchy (product hazards), 763–764
- Design of experiments (DOE), 401, 404
 - DMADV, 661, 662
 - DMAIC, 644–645, 647–649
 - generating input data, 411–412
- Design patents, 774
- Design phase (DMADV), 661–663
- Design scorecard, 660
- Desktop manufacturing, 180
- Deterministic inventory models, 65
- Diagonal pitch (joints), 262
- Diamonds, 127, 154
- Diamondlike carbon (DLC) films, 247, 249
- DIBS (dual-ion-beam sputtering), 248
- Die casting, 37–38, 223
- Digital (term), 807
- Digital business, 20
- Digital journals, 820
- Digital repositories, 820, 821
- Digitizing, programming by, 347
- Dimension (statistical quality control), 325
- Dinking, 215
- Direct infringement (patents), 797
- Direct numerical control (DNC) systems, 351, 352
- Disaggregation levels (production planning), 76
- Discounted inventory model, 71
- Dispatching rules (job sequencing/scheduling), 94–97
- Distinctness requirement (patents), 786
- Distributed artificial intelligence (DAI), 521
- Distributed systems, 517–520
- Distribution logistics, 108–109
- Divisional patent applications, 794–795
- DLC (diamondlike carbon) films, 247, 249
- DMADV (TQM), 636–637, 656–664
 - analyze phase, 660–661
 - define phase, 656–657
 - design phase, 661–663
 - measure phase, 657–660
 - verification and validation phase, 663–664
- DMAIC (DMAIC, TQM), 621, 636–656
 - analyze phase, 643–649
 - control phase, 653–656
 - define phase, 637–639
 - improve/innovate phase, 649–653
 - measure phase, 639–643
- DMP Tool, 820–821
- DNC (direct numerical control) systems, 351, 352
- Documentation:
 - in assessment for ECM, 189–190
 - of design process, 770–771
 - in DMAIC, 650, 652
 - original patent documents, 796
 - process, 664
- Documentation tree, 770
- Document delivery (online information), 817–818
- DoD, *see* U.S. Department of Defense
- DOE, *see* Design of experiments
- Double acceptance sampling, 335–336
- Double declining balance (DDB) depreciation, 600
- Double-patenting, 794
- Downloading, 807
- Drawing (metals):
 - cold-working processes, 208, 215–217
 - environmentally benign manufacturing, 41
 - estimating load, 204, 205
 - hot-working processes, 203–205
- Drilling:
 - for bolts and rivets, 261–262
 - tool wear factors, 122
- Drilling machines, 133–140
 - accuracy of, 138–140
 - classification of, 140
- Driver training, 735–736
- Drucker, Peter, 9, 13
- Dry sand molds, 218
- DSL (digital subscriber line), 807
- DSpace, 807, 820
- Dual-ion-beam sputtering (DIBS), 248
- Duty of candor (patents), 791
- DVD (digital video disc), 807
- Dynamic seals, 296–319
 - brush seals, 313–319
 - bristle stiffening, 316
 - brush pack considerations, 315
 - design considerations, 314–315

- Dynamic seals (*continued*)
 - flow modeling, 317–318
 - implementation issues, 315
 - leakage performance, 316–317
 - materials for, 318–319
 - pressure closing, 316
 - seal hysteresis, 315, 316
 - for emissions control, 301–305
 - application guide, 304, 305
 - double seals, 303–304
 - sealing approaches for, 302–304
 - single seals, 302
 - tandem seals, 302, 303
 - honeycomb seals, 312–313
 - initial selection of, 296–298
 - labyrinth seals, 308–312
 - applications, 310, 311
 - computer analysis tools for, 311
 - configurations, 309–311
 - leakage flow modeling, 309, 310, 312
 - and rotordynamic stability, 311
 - mechanical face seals, 296, 298–301
 - emissions performance of, 301
 - leakage, 299–301
 - materials selection for, 300–301
 - seal balance, 296, 298–299
 - seal face flatness, 300
 - noncontacting seals for high-speed/aerospace
 - applications, 304–308
 - radial lip seals, 296
 - turbine engine seals, 296
 - Dynamic structures, 519
- E**
- EAs (environmental assessments), 694
 - EBM (electron beam machining), 172–173
 - E-business, 20
 - ECD (electrochemical deburring), 166
 - ECDG (electrochemical discharge grinding), 166, 167
 - ECH (electrochemical honing), 167–168
 - EC inspection, *see* Eddy current inspection
 - ECM, *see* Environmentally conscious manufacturing
 - ECM (electrochemical machining), 168–169
 - E-commerce, 396
 - Economics, *see* Engineering Economics
 - ECP (electrochemical polishing), 169, 170
 - ECS (electrochemical sharpening), 169, 170
 - ECT (electrochemical turning), 170, 171
 - Eddy current (EC) inspection, 469–475
 - capabilities, 444
 - impedance plane, 470–473
 - lift-off of inspection coil from specimen, 472–474
 - probes and sensors, 474
 - skin effect, 470
 - EDG (electrical discharge grinding), 173
 - Edge bending, 211
 - EDM, *see* Electrical discharge machining
 - EDS (electrical discharge sawing), 174
 - Education programs (manufacturing systems), 22–23
 - EDWC (electrical discharge wire cutting), 174–175
 - Effective interest rate, 585, 587–588
 - EFQM, 685
 - eFunda, 817
 - Eight laws of engineered systems evolution, 363
 - EISs (environmental impact statements), 694
 - E-kanban systems, 624
 - Elastic limit (bolts), 270
 - Elastic region (fasteners), 270
 - Elastomer properties, 286
 - Electrical discharge grinding (EDG), 173
 - Electrical discharge machining (EDM), 32, 47, 173, 174
 - Electrical discharge sawing (EDS), 174
 - Electrical discharge wire cutting (EDWC), 174–175
 - Electrochemical deburring (ECD), 166
 - Electrochemical discharge grinding (ECDG), 166, 167
 - Electrochemical grinding, 167
 - Electrochemical honing (ECH), 167–168
 - Electrochemical machining (ECM), 168–169
 - Electrochemical polishing (ECP), 169, 170
 - Electrochemical sharpening (ECS), 169, 170
 - Electrochemical turning (ECT), 170, 171
 - Electroforming (plastic injection molding), 47
 - Electromechanical machining (EMM), 160
 - Electron beam evaporation, 239–240
 - Electron beam machining (EBM), 172–173
 - Electron beam welding, 44
 - Electroplating, 232
 - Electropolishing (ELP), 158, 178, 179, 230
 - Electrostream (ES), 170, 171
 - ELP, *see* Electropolishing
 - EMAS (environmental management scheme, European Union), 677
 - Embossing (metals), 217
 - Emissions control seals, 301–305
 - application guide, 304, 305
 - double seals, 303–304
 - sealing approaches for, 302–304

- single seals, 302
- tandem seals, 302, 303
- EMM (electromechanical machining), 160
- Employee liability, 751–754
 - agency and authority, 752–753
 - employment agreements, 753–754
 - intellectual property, 754
 - negligence and standard of care, 751–752
- Employment agreements, 753–754
- Empowerment, workforce, 15–16
- EMS (Environmental Management System)
 - registrars, 670
- Enablement requirement (patents), 785–786
- Enamels, 231
- Encyclopedias, 825–826
 - Encyclopedia of Polymer Science and Engineering*, 825
- End cutting edge angle, 127
- Energy expenditure, for transport, 31
- Energy Management Handbook*, 826
- Energy sources/forms, environmental damage
 - from, 185
- Energy transfer analysis, 701
- Engineering controls (safety):
 - alternatives to, 715, 717–719
 - for machine tools, 710–713
 - and PPE, 715
- Engineering Economics, 581–603
 - after tax analysis, 601
 - break-even analysis, 602
 - capitalized cost, 599
 - capital recovery, 598
 - cash flows and time value of money, 583–584
 - comparing alternatives:
 - defining alternatives, 591–592
 - through figures of merit, 592–598
 - depreciation, 599–600
 - equivalence, 584–588
 - ethics in, 582–583
 - inflation, 600–601
 - manipulation of cash flows, 588–591
 - replacement studies, 599
 - risk analysis, 602
 - sensitivity analysis, 601–602
 - types of decisions in, 583
- Engineering Index*, 823
- Engineering societies, as information source, 827–828
- Engineering Village, 816
- Engnet, 816
- Enterprise integration, 519
- Enterprise resource planning (ERP), 86
- Environmental assessments (EAs), 694
- Environmental impact statements (EISs), 694
- Environmentally benign manufacturing, 14, 29–50
 - machining processes, 31–32
 - manufactured product, 50
 - manufacturing and supply chain, 30–31
 - manufacturing processes, 31–50
 - machining, 31–32
 - metal casting, 32–38
 - metal-forming, 38–41
 - metal-joining, 41–45
 - plastic injection molding, 45–50
 - metal casting processes, 32–38
 - metal-forming processes, 38–41
 - metal-joining processes, 41–45
 - plastic injection molding processes, 45–50
- Environmentally conscious manufacturing (ECM), 17–19
 - assessment and improvements for achieving, 189–193
 - challenge for, 29
 - components of, 183–184
 - system effects on, 187–188
 - transport in, 31
- Environmental management scheme (EMAS, European Union), 677
- Environmental Management System (EMS)
 - registrars, 670
- Environmental management systems standard (ISO 14000), 675–677
- Environmental risk training, 736–739
- Environment consciousness (material handling), 503
- EPA, *see* U.S. Environmental Protection Agency
- EPC (European Patent Convention), 803
- EPO (European Patent Office), 803
- Equivalence (monetary transactions), 584–588
- Ergonomics, 709. *See also* Human factors
 - engineering (ergonomics)
- Ergonomics principle (material handling), 499
- ERP (enterprise resource planning), 86
- Error analysis, 54, 61, 63
- ES (electrostream), 170, 171
- eStandards Store, 818
- Etching (plastic injection molding), 47
- Ethics, 582–583
- European Excellence Award, 685–686
- European Patent Convention (EPC), 803
- European Patent Office (EPO), 803
- European Quality Award, 678, 687
- European Union environmental management
 - scheme, 677

- Evaluation:
 - in assessment, 189
 - of manufacturing systems, *see* Assessment of manufacturing systems
 - of projects, *see* Project evaluation and selection
 - Evaporative processes (PVD), 239–242
 - Expandable-bead molding (plastics), 225
 - Ex parte* Jepson, 787
 - Ex parte* reexamination of patent validity, 796–797
 - Expected outcome approach, 701
 - Expertise:
 - as basis of interpersonal power, 569
 - technical, building, 575
 - Exponential smoothing, 54, 57–58
 - Express warranty, 759
 - Extended products, 19
 - External business environment, 564
 - Extrusion:
 - environmentally benign manufacturing, 40
 - metals, 201–202, 208–210
 - plastics, 225
- F**
- Fabricators, 341
 - Failure mode and effect (FME), 700
 - Failure stress (bolts), 270
 - FAQ (frequently asked questions), 807
 - Fastening, 255–256. *See also* Mechanical fasteners
 - three-dimensional strategy, 256–259
 - three-part assemblies, 260–261
 - Fatigue failure (mechanical fasteners), 280–282
 - Fault tolerance, 519
 - Fault tree analysis, 701–703
 - Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA, 1996), 697
 - Federal Water Pollution Control Act Amendments (1972), 694
 - Federated searching, 807
 - Federations (agents), 530
 - Fedora, 807, 820
 - Fee-based database services, 815–816
 - Feed(s):
 - abrasive machining, 154
 - metal cutting, 129
 - milling, 140, 141
 - FIFRA (Federal Insecticide, Fungicide, and Rodenticide Act, 1996), 697
 - File server, 807
 - Fillet(s) (bolts), 281
 - Film-based radiography, 452–453
 - Filmless radiography (FR), 454, 455
 - Financial rewards, as motivator, 570
 - Fine William T., 707
 - Finishing, 153–158
 - abrasives, 154–156
 - barrel finishing, 229
 - belt finishing, 230
 - gears, 146
 - surface, 157–158, 227
 - temperature, 156–158
 - FIPA (Foundation for Intelligent Physical Agents), 528
 - Fire protection training, 739
 - Firewalls, 807, 811, 812
 - First-aid training, 739
 - Fishbone diagrams, 15, 643–644
 - Fisher, Sir Ronald, 648
 - Fit dependency, 526
 - 5S, 621–622, 730–733
 - Fixed-position layout (production plants), 342
 - Flammability hazards, 721–722
 - Flanges (seal technology), 288
 - Flange rotation (bolts), 271
 - Flanging (metals), 213
 - Flattening (metals), 213
 - Flexible manufacturing systems (FMS):
 - automation technologies in, 352–353
 - coordination in, 526
 - hierarchical computer control, 351
 - intelligent control, 516–517
 - multiobjective model for operation allocation/selection, 506–508
 - Floating agents, 552–554
 - Flood Control Act of 1936, 597
 - Florida Power and Light, 680
 - Flow dependency, 526
 - Flow shops, 90–92
 - FME (failure mode and effect), 700
 - FMS, *see* Flexible manufacturing systems
 - Food safety standard (ISO 22000), 677–678
 - Ford, Henry, 6, 20
 - Ford Motor Company, 7, 392, 516, 618
 - Forecast (term), 54
 - Forecasting, 54–63
 - causal methods, 58–60
 - definitions related to, 54–55
 - error analysis, 61, 63
 - qualitative, 55–58
 - time series analysis, 60–62
 - Foreign patents, *see* International/foreign patents
 - Forged-plastic parts, 226
 - Forging (metals):
 - environmentally benign manufacturing, 40
 - hot-working processes, 199–201
 - Form cutting (gears), 144–146

- Foundation for Intelligent Physical Agents (FIPA), 528
- Foundry industry, 36
- Four-Stage Team Development Model, 572–573
- FR (filmless radiography), 454, 455
- Fractional factorial experiments, 648
- Free database services, 816–817
- Freese and Nichols, 636
- Frequently asked questions (FAQ), 807
- Friction-type connections, 262, 268–270
- Frit, 231
- From the American System to Mass Production 1800–1932* (D. A. Hounshell), 6
- Full factorial experiments, 648
- Full-mold casting, 38, 224
- Future value (FV), 592–595
- ## G
- Gas-flame welding, 41, 42
- Gaskets, 283–289
 - bolt pattern, 288–289
 - elastomer properties, 286
 - elevated-temperature service, 289
 - flange surfaces, 288
 - flange thickness, 288
 - gasket factors, 284–285
 - materials, 283–287
 - metallic, 285, 287
 - minimum design seating stress, 284–285
 - nonmetallic, 285
 - required bolt load (ASME method), 287
 - thickness and compressibility, 289
- Gasket crush (bolts), 271
- Gates (injection molding), 47
- GE (General Electric), 319
- Gear manufacturing, 143–146
- Gemba* walks, 624
- Genchi genbutsu* (go and see), 624
- General Electric (GE), 319
- Generalized partial global planning (GPGP), 527
- General Motors Saturn Project, 17
- Generating process (gears), 145
- Geometric gradient (cash flows), 588
- Gerstner, Lou, 17
- Global commons, 14
- Global rules (job sequencing/scheduling), 94
- Glow discharge plasma (PVD), 236–238
- GNH (gross national happiness), 25
- Goals, 566, 576
- Goal-chasing methods, 102
- Goals-coordinating method, 102
- Goal programming, 74, 75
- Google Scholar, 817
- Governments:
 - patent infringement by, 800
 - publication resources from, 827
- Government regulatory safety requirements, 693–700
 - Environmental Protection Agency, 694–697
 - Occupational Safety and Health Administration, 697–700
 - state-operated compliance programs, 698–700
- GPGP (generalized partial global planning), 527
- Gradient series (cash flows), 588
- Graham v. John Deere*, 783
- Graphical data analysis, 644
- Graphical user interface (GUI), 807
- G* ratio (grinding ratio), 155
- Green sand molds, 218
- Grinding, 153–158
 - abrasives, 154–156
 - electrical discharge, 173
 - electrochemical, 167
 - electrochemical discharge, 166, 167
 - low-stress, 159, 160, 163
 - temperature, 156–158
- Grinding fluids, 157
- Grinding machines, 157
- Grinding ratio (*G* ratio), 155
- Gross-hazard analysis, 700
- Gross national happiness (GNH), 25
- Gross requirements, 77–81
- Group technology, 353–358
 - computer-aided process planning, 358
 - machine cell designs, 357–358
 - part family formation, 354
 - parts classification and coding, 354–357
 - production flow analysis, 357
- GUI (graphical user interface), 807
- Guilds, 5
- Gun drills, 139
- Guterl, F., 9
- ## H
- HACCP (hazard analysis and critical control points), 677–678
- Handbooks, 825–826
- Handbook of Energy Efficiency and Renewable Energy*, 826
- Handbook of Human Factors and Ergonomics*, 826
- Hardware, safety design for, 719, 720
- Harley-Davidson, 22
- Hazards:
 - after-market product hazards, 768–770
 - safety issues, *see* Safety engineering

- Hazard analysis (product defects), 762–763
- Hazard analysis and critical control points (HACCP), 677–678
- Hazard Communication Standard (HCS), 723
- Hazard criticality ranking, 701
- Hazard index (product defects), 763
- Hazardous and Solid Waste Amendments (HSWA), 1984), 695
- Hazardous materials, 184, 185
 - defined, 184
 - HAZWOPER training, 740
 - National Emission Standards for Hazardous Air Pollutants, 301. *See also* Emissions control
 - NFPA “hazard diamond,” 720–722
 - Toxic Substances Control Act, 695
- Hazardous material classification system, 720–722
- Hazardous mechanical motions, 711, 712.
 - See also* Machine safeguarding methods
- Hazardous waste, 695
 - Hazardous and Solid Waste Amendments, 695
 - HAZWOPER training, 740
 - Superfund Act (CERCLA), 695–696
 - Superfund Amendments and Reauthorization Act, 696
- Hazardous Waste Operations and Emergency Response Standard (HAZWOPER), 740
- Hazard Ranking System (HRS), 696
- Hazard recognition training, 739–740
- HAZWOPER (Hazardous Waste Operations and Emergency Response Standard), 740
- HAZWOPER training, 740
- HCS (Hazard Communication Standard), 723
- HDM (hydrodynamic machining), 159, 163
- Health hazards, 720, 721. *See also* Safety engineering
- Heat Index, 736–738
- HERF (high-energy-rate forming), 217
- Heuristics (job sequencing/scheduling), 94–97
- Hewlett-Packard (HP), 21, 624
- Hierarchical computer control, 351–352
- Hierarchical model (agents), 529
- Hierarchical structure, 7, 8
- Hierarchies (agents), 529
- High-carbon steels, 126
- High-energy-rate forming (HERF), 217
- High-performance teams, 572–574
- High-power impulse magnetron sputtering (HIPIMS, HPPMS), 244
- High-speed applications (seals):
 - brush seals, 315, 317
 - noncontacting, 304–308
- High-speed steels (HSS), 126, 131
- High-temperature service (seals), 319
 - gaskets, 289
 - O-rings, 291
 - packings and braided rope seals, 293–295
- HIPIMS (high-power impulse magnetron sputtering), 244
- Histograms, 626–627, 644
- HMS (holonic manufacturing systems), 518
- Hobbing (metals), 210
- Hoffman, Bryce, 7
- Hoists, 505
- Holarchies (agents), 529
- Holding costs, 64
- Holons, 518
- Holonic agents, 518
- Holonic manufacturing systems (HMS), 518, 520
- Honeycomb seals, 312–313, 317
- Honing, 157–158, 167–168
- HOQ (House of Quality), 15, 548
- Horizon, 54
- Host (computer network), 807
- Hot-chamber die casting, 37, 223
- Hot-dip plating, 232
- Hot runner molding, 47
- Hot-working processes (metals), 38, 195–207
 - classification of, 197
 - drawing, 203–205
 - environmentally benign manufacturing, 38, 39
 - extrusion, 201–202
 - forging, 199–201
 - piercing, 207
 - pipe welding, 206–207
 - rolling, 197–199
 - spinning, 206
- Hounshell, D. A., 6
- House of Quality (HOQ), 15, 548
- HP (Hewlett-Packard), 21, 624
- HPPMS (high-power impulse magnetron sputtering), 244
- HRS (Hazard Ranking System), 696
- HSS (high-speed steels), 126, 131
- HSWA (Hazardous and Solid Waste Amendments, 1984), 695
- HTML (HyperText Markup Language), 807
- Human error analysis, 701
- Human Factors and Ergonomics Design Handbook*, 826
- Human factors engineering (ergonomics), 708–711
 - general population expectations, 710, 711
 - human-machine relationships, 708–709
 - in material handling, 499
 - principles of, 709–710
- Human-machine relationships, 708–709
- Human relations movement, 571

- Hydra, 820
 Hydrodynamic face seals, 304, 305, 307
 Hydrodynamic machining (HDM), 159, 163
 Hydroform process, 217
 Hydrostatic extrusion, 210
 Hydrostatic face seals, 304, 306
 HyperText Markup Language (HTML), 807
 Hypothesis testing, 644
- I**
- IAC (ion-assisted coating), 247
 IAQG (International Aerospace Quality Group), 672
 IBAD (ion-beam-assisted deposition), 247, 248
 IBM, 17, 20
 Ideas, patentable inventions vs., 777–778
 Ideal final result (IFR), 364, 365
 Ideality principle (TRIZ), 364–365
 Identity theft, 808, 812
 IEC (International Electrotechnical Commission), 669
 IFCs (industry foundation classes), 395
 IFR (ideal final result), 364, 365
 IHS (Information Handling Services), 827
 Impact extrusion, 209
 Impact machining, 159
 Impression-die drop forging (metals), 200, 201
 Improve/innovate phase (DMAIC), 649–653
 Improvement initiatives, 4, 14–15
 IMS (intelligent manufacturing systems) program, 518, 520
 Indexes, 824–825
 Index numbers, 55, 60
 Inducement contribution model, 566
 Inducement of infringement (patents), 797
 Industrial engineering function, 341
 Industrial Revolution, 5
 Industrial robots, 349–351
 applications, 350–351
 configurations, 349, 350
 control and programming, 349, 350
 defined, 349
 in flexible manufacturing systems, 353
 in jobs with safety risks, 720
 Industry foundation classes (IFCs), 395
 Industry-specific quality awards, 686
Industry Week America's Best Plants Award, 686, 687
 Inflation, 600–601
 Inflation rate, 601
 Information Handling Services (IHS), 827
Information Industry Market Place, 825
 Information sources/resources, 823–828
 codes, 826
 for education programs, 23
 encyclopedias and handbooks, 825–826
 engineering societies, 827–828
 government publications, 827
 indexes and abstracts, 824–825
 libraries, 828
 for nondestructive inspection, 442–443
 online, 805–821
 access options, 809–811
 database services, 813–817
 document delivery options, 817–818
 future of online access, 819–821
 indexes and abstracts, 825
 managing online literature, 819
 options for using, 821
 security issues, 811–812
 terms related to, 806–809
 primary literature, 823–824
 specifications, 826
 standards, 826–827
 Information technology, 519–520
 Infotrieve, 818
 Infrared (IR) cameras, 468–469
 IngentaConnect, 818
 Injection-molded carbon fiber composites, 226
 Injection molding (plastics), 45–50, 224
 Innovation, *see* TRIZ
 Inputs, 398
 In-scope/out-of-scope, 643
In Search of Excellence (Tom Peters and Robert H. Waterman), 624
 Inspection:
 in assessment of systems, 191
 nondestructive, *see* Nondestructive inspection (NDI)
 Instability hazards, 722
 Institutional repository (IR), 807, 808
 Instructions (product defects), 761–762
 Insurance, 756
 Intellectual property, 753–754. *See also* Patents
 Intelligent agents, 521–522
 Intelligent control of MHSs, 515–554
 agent technology, 520–531
 agent architectures, 523–526
 interagent communication, 526–528
 multiagent systems, 520–523
 organization of agents, 528–531
 types of agents, 523
 baggage handling system case study, 531–544
 agent design, 534
 agent interactions and ontology, 538–540
 internal agent reasoning, 540–543

- Intelligent control of MHSs (*continued*)
 - performance criteria, 532–533
 - toploader, 534–538
 - worst-case scenario, 533–534
- distributed systems, 517–519
- flexible manufacturing, 516–517
- history, 516
- improving results of, 550–554
 - active, sleeping, and locked agents, 550–552
 - floating agents, 552–554
 - predecessor validation, 552, 553
- material handling in anodization system case study, 543–550
 - agent design, 546–548
 - interactions, 548–550
 - PACO approach, 545–546
- new challenges for, 519–520
- Intelligent manufacturing systems (IMS) program, 518, 520
- Interest rates, 585, 587–588
- Interferences (patent disputes), 800–801
- Intermediate pitch (joints), 262
- Internal rate of return (IRR), 595–597
- Internal threads, cutting and forming, 146–147
- International Aerospace Quality Group (IAQG), 672
- International Automotive Task Force, 673
- International Electrotechnical Commission (IEC), 669
- International Federation of the National Standardizing Associations (ISA), 669
- International/foreign patents, 776, 779, 801–803
- International Organization for Standardization (ISO), 183, 667, 669. *See also* ISO standards
- International quality standards, 669. *See also* Quality management systems
- Internet, 805–806, 808
- Interoperability, 519
- Inter partes* reexamination of patent validity, 796, 797
- Interpersonal power, 569
- InteRRaP agent architecture, 524–526
- Interrelationship digraph, 629, 630
- Intranet, 808, 811
- Inventions, 777–778
- Inventiveness, levels of, 363–364
- Inventive problems, typical problems vs., 370
- Inventive solutions, *see* TRIZ
- Inventory, 65
- Inventory control function, 342
- Inventory costs, 74
- Inventory models, 63–73
 - definitions related to, 65–66
 - modeling approach, 66–73
 - types of, 65
- Inventory unit, 77
- Inverse storage (warehousing), 510
- Investment casting, 224
- Investment process, 224
- Ion-assisted coating (IAC), 247
- Ion-beam-assisted deposition (IBAD), 247, 248
- Ion beam processes (PVD), 247–248
- Ionization enhancement (PVD), 240, 241
- Ion plating, 237
- IR (institutional repository), 807, 808
- IR (infrared) cameras, 468–469
- Ironing (metals), 217
- IRR (internal rate of return), 595–597
- Irregular movements, in time series, 60
- ISA (International Federation of the National Standardizing Associations), 669
- Ishikawa diagrams, 15, 643–644
- Islandora, 820
- ISO, *see* International Organization for Standardization
- Isolation, as safety control, 718, 719
- ISO standards:
 - accreditation bodies, 669
 - ISO 9000: quality management systems—fundamentals and vocabulary, 668–670
 - ISO 9001: quality management systems—requirements, 669–673
 - for automotive production and relevant service part organizations (2008), 672, 673
 - and Baldrige Core Values, 670
 - certification/registration, 670, 671
 - and ISO/TS 16949, 672, 673
 - Section 7, product realization, 669
 - ISO 9004: quality management systems—managing for sustained success of an organization, 670
 - ISO 13485: medical devices, 673–674
 - ISO 14000: environmental management systems, 675–677
 - ISO 14001: environmental management system registration, 183, 675–677
 - ISO 19011: guidelines for quality and/or environmental management systems auditing, 670, 675
 - ISO 22000: food safety management, 677–678
 - ISO/TS 16949: automotive production/service part organizations, 672, 673
- Isothermal rolling, 38, 40, 197–199

ISO/TS 16949: automotive production/service part organizations, 672, 673
 Item cost, 64
 I-TRIZ, 371

J

Japanese manufacturing philosophy, 104–107
 just-in-time/kanban, 104–107
 time-based competition, 107
 Japanese Union of Scientists and Engineers (JUSE), 678, 679
 Japan Quality Award (JQA), 686, 687
 Jepson-style patent claims, 787
 Jibs, 505
 JIT, *see* Just-in-time
 Job hazard analysis training, 742–744
 Job sequencing and scheduling:
 assembly line balancing, 97–104
 definitions related to, 97–98
 design of assembly line, 100
 mixed-model assembly lines, 101–103
 parallel line balancing, 103–104
 structure of problem, 99
 techniques for, 100–101
 flow shops, 90–92
 heuristics/priority dispatching rules, 94–97
 job shops, 92–94
 single-machine problem, 88–90
 structure of sequencing problem, 87–88
 Job shops, 92–94, 342
 more than two machines and n jobs, 93–94
 two machines and n jobs, 92–93
 Johnson's sequencing algorithm, 91–92
 Joining, 255–256. *See also* Mechanical fasteners
 Joints, bolted and riveted:
 clamping force upper limits, 271–272
 complex butt joint (bearing-type connection), 266–268
 efficiency, 264
 friction-type connections, 268–270
 simple lap joint strength, 264–265
 slip characteristics evaluation, 276–277
 theoretical behavior of joint under tensile loads, 272–276
 types, 262–264
 Joint and severable liability, 752
 Joint inventorship, 779–780
 JQA (Japan Quality Award), 686, 687
The Jungle (Upton Sinclair), 6
 JUSE (Japanese Union of Scientists and Engineers), 678, 679

Just-in-time (JIT), 21, 104–107, 509
 in lean, 623
 mixed-model assembly lines, 102, 103
 support for, 20
 warehousing, 509

K

Kaizen, 21, 620
 Kaizen events/blitzes, 621, 622
 Kanban, 104–107, 109–110, 623–624
 Kano, Noriaki, 657
 Kano model, 657, 658
 Kellogg Company, 25
Keweenaw Oil v. Bicron Corp., 778
 Key learnings (DMADV), 664
 Keynes, John Maynard, 25
 Kidder, T., 9
Kirk–Othmer Encyclopedia of Chemical Technology, 825
 Knovel, 816
 Knowledge Query and Manipulation Language (KQML), 527
 Known demand (inventory), 64
 Kondratieff waves, 8

L

Labor costs, 16
 Labyrinth seals, 308–312
 applications, 310, 311
 brush seals vs., 317
 computer analysis tools for, 311
 configurations, 309–311
 leakage flow modeling, 309, 310, 312
 and rotordynamic stability, 311
 Lacquers, 231
 Lancing (metals), 215
 Lap joints, 262, 264–265
 Lapping, 158
 Lap welding, 207
 Laser beam machining (LBM), 175–176
 Laser beam torch (LBT), 176–177
 Lathe size, 132
 Laws of systems evolution, 374–375
 Lawsuits, 751
 LBM (laser beam machining), 175–176
 LBT (laser beam torch), 176–177
 LCM (life-cycle management), 14
 LDR (linear decision rule) technique, 74
 Leadership, 577
 adapted to situations, 574
 of engineering teams, 570–573

- Leadership (*continued*)
 - evolution of, 6–8
 - of high-performance teams, 573
 - in ideal manufacturing system, 24
 - and lean management, 633, 634
 - managerial, 563
 - of project evaluation team, 615
- Lead time, 65, 77
- Leakage:
 - brush seals, 316–317
 - labyrinth seals, 309, 310, 312
 - mechanical face seals, 299–301
- Lean management, 21, 617–634
 - history of, 617–618
 - leadership and, 633, 634
 - lean tools, 621–624
 - 5S, 621–622
 - gemba* walks, 624
 - genchi genbutsu* (go and see), 624
 - just-in-time manufacturing, 623
 - kanban, 623–624
 - poka-yoke, 623
 - takt time, 624
 - total productive maintenance, 623
 - visual factory, 622
 - management tools, 629–634
 - activity network diagram, 631, 634
 - affinity diagram, 629, 630
 - interrelationship digraph, 629, 630
 - matrix diagram, 631, 633
 - prioritization matrix, 629, 631, 632
 - process decision program chart, 631, 633
 - tree diagram, 629, 631
 - philosophy and deployment:
 - Kaizen, 620
 - Kaizen event/blitz, 621, 622
 - muda* (waste), 619–620
 - plan–do–check–act cycle, 620–621
 - value stream, 618–619
 - traditional quality control tools, 624–629
 - cause-and-effect diagram, 625
 - check sheets, 625, 626
 - control charts, 625, 626
 - histograms, 626–627
 - Pareto charts, 627, 629
 - scatter diagrams, 628, 629
 - stratification, 628, 629
- Lean manufacturing, 618
- Lean production, 516
- Lean Six Sigma, 636. *See also* Total quality management (TQM)
- Least cost problems, 583
- Legal issues in manufacturing:
 - enforcing against patent infringement, 797–801
 - responsibility for products through life cycle, 510
- Legal requirements for engineers, 749–771
 - and art of engineering, 749–751
 - design process documentation, 770–771
 - product liability, 757–770
 - after-market hazards, 768–770
 - assumption of risk, 767–768
 - contributory/comparative negligence, 766
 - corrective actions, 769–770
 - defense to, 764–768
 - design flaws, 760–761
 - express warranty and misrepresentation, 759
 - hazard analysis, 762–763
 - hazard index, 763
 - instructions and warnings, 761–762
 - laws of, 757–759
 - nature of product defects, 760–762
 - production or manufacturing flaws, 760
 - recalls, retrofits, and continuing duty to warn, 768–770
 - safety standards, 765–766
 - state of the art defense, 764–765
 - strict liability, 758–759
 - uncovering product defects, 762–764
- professional liability, 751–757
 - agency and authority, 752–753
 - business liability, 755–757
 - case study, 757
 - contractual obligations, 755–756
 - employee liability, 751–754
 - employment agreements, 753–754
 - insurance for individual engineers, 756
 - intellectual property, 754. *See also* Patents
 - negligence and standard of care, 751–752
 - negligence for services, 755
- Lenovo, 21
- Levels of manufacturing systems, 185–186
- Liability, *see* Product liability; Professional liability
- Liberty Brass, 22
- Libraries, as information source, 828
- Life cycle:
 - of improvement initiatives, 4
 - of plastic products, 49–50
 - product, 392, 393
- Life-cycle field (in LCM), 14
- Life-cycle management (LCM), 14
- Life cycle principle (material handling), 503
- Linear decision rule (LDR) technique, 74
- Linear programming, 74
- Liquid baths (cleaning), 228–229

Liquid honing, 229
 Liquid penetrants (NDI), 444–447
 Listserv, 808
 Load excursions, mechanical fasteners and, 282
 Local governments, patent infringement by, 800
 Local rules (job sequencing/scheduling), 94
 Locators (fastening), 256–258
 Locks (fastening), 259
 Locked agents, 550–552
 Lockout box training, 740–741
 Logistics, 13, 108–109
 Long pitch (joints), 262
 Lost-wax casting, 37, 224
 Lot size models, 74, 75
 Lot sizing techniques, 85–86
 Low-stress grinding (LSG), 159, 160, 163
 Lumpy demand, 64, 66
 Lumpy demand inventory models, 65

M

McAfee, 812
 Machinability, 128–129
 Machinability data prediction, 347
 Machine (legal term), 778
 Machine cell designs, 357–358
 Machine guard training, 741
 Machine safeguarding methods, 714–717
 devices, 716–717
 feeding and ejection methods, 717
 general classification of, 714
 guards, 715
 training in use of, 741
 Machine tools engineering controls, 710–713
 Machining:
 abrasive, 153–158
 abrasives, 154–156
 temperature, 156–158
 abrasive flow, 158–159, 162
 abrasive jet, 159, 162
 chemical, 178
 combustion, 179
 conventional processes, 117
 electrical discharge, 173, 174
 electrochemical, 168–169
 electromechanical, 160
 electron beam, 172–173
 environmentally benign manufacturing, 31–32
 gears, 144–146
 hydrodynamic, 159, 163
 laser beam, 175–176
 photochemical, 178–179
 plasma beam, 177, 178
 plastics, 153
 sand casting allowances, 219
 shaped-tube electrolytic machining, 171–172
 thermally assisted, 160, 164
 thermochemical, 179–180
 total form, 161, 164
 ultrasonic, 161, 163, 165
 water-jet, 163–165
 Machining power, 119–121
 Macrolevel (process planning), 188
 MACRS, 600
 Magnetic particle inspection, 444, 465–468
 Magnetron sputtering (PVD), 242, 245–247
 high-power impulse, 244
 medium-frequency ac, 243–244
 Maintenance hazard analysis, 701
 Makespan, 91, 93
 Make-to-order, 187
 Make-to-stock, 187
 Malcolm Baldrige Awards, 10
 criteria for performance excellence, 11
 cycle time for Baldrige-based awards, 678
 National Quality Award, 668, 670, 678,
 680–684, 687
 Performance Excellence Award, 635–636
 and TQM applications, 636
 Malpractice insurance, 756
 Malware, 808
 Management. *See also specific topics*
 core issues/challenges in, 560
 evolution of, 6–8
 in ideal manufacturing system, 24
 improvement initiatives, 4
 of people, 559–577
 critical issues/challenges in, 559–561
 engineering personnel, 561–564
 engineering teams, 570–573
 motivation, 564–569
 power profile in, 569–570
 recommendations for, 573–577
 of project evaluation team, 615
 of safety function, 727–735
 accident prevention, 727–729
 eliminating unsafe conditions, 729–734
 oversight, 744–745
 principles for, 728, 730
 supervisor's role, 727
 unsafe conditions checklist for mechanical or
 physical facilities, 734–735
 workforce considerations, 15–17
 Management coefficient models (MCM), 74, 75
 Management oversight and risk tree (MORT), 700
 Management process, unified, 575–576
 Managerial leadership, 563

- Manual spinning, 206
- Manufacturers (as legal term), 778
- Manufacturing cells, 351, 352
- Manufacturing cycle, 341
- Manufacturing engineering function, 341
- Manufacturing flaws, 760
- Manufacturing function, 342
- Manufacturing model with shortage permitted (inventory), 69, 70
- Manufacturing model with shortage prohibited (inventory), 69
- Manufacturing process flow diagrams, 398
- Manufacturing resources planning (MRP-II), 77, 86
- Manufacturing systems, 3–26, 183–193
 - assessment, 189–193
 - data collection, 190–191
 - planning assessments, 189–190
 - reporting and project formulation, 192–193
 - site visit and inspection, 191
 - components of, 12–13
 - defined, 17
 - environmentally conscious, 17–19. *See also* Environmentally benign manufacturing
 - components of, 183–184
 - system effects on, 187–188
 - evolution of leadership and management in, 6–8
 - history of, 4–5
 - ideal system of the future, 24–26
 - implementation of, 20–24
 - education programs, 22–23
 - measuring results, 23–24
 - real-world examples, 20–22
 - vertical integration, 20
 - levels of systems, 185–186
 - measurement and organization of, 10–12
 - plan–do–check–act cycle, 187
 - and quest for stability, 8–10
 - vision for improvement/reconfiguration of, 13–15
 - workforce social engineering, 15–17
- MAPs (mission assurance provisions), 668, 674–675
- Markets, as agents, 530
- MARR, *see* Minimal attractive rate of return
- MASs (multiagent-based systems), 520–523. *See also* Agents
- Mass customization, 515, 516
- Massive open online course (MOOC), 808, 820
- Mass production, 342
- Master production schedule, 78
- Master scheduling, *see* Aggregate planning (AP)
- Material characteristics, 400
- Material handling devices (MHDs), 504–506
 - automated guided vehicles, 505
 - autonomous vehicle storage and retrieval systems, 505
 - choosing, 506
 - conveyors, 504
 - hoists, cranes, jibs, 505
 - palletizers, 504
 - robots, 505
 - trucks, 504
 - warehouse MHDs, 505
- Material handling equation, 506
- Material handling function, 342
- Material handling plan, 498
- Material handling systems (MHSs), 497–511
 - anodization system case study, 543–550
 - agent design, 546–548
 - interactions, 548–550
 - PACO approach, 545–546
 - AVS/RS case study, 510–511
 - baggage handling system case study, 531–544
 - agent design, 534
 - agent interactions and ontology, 538–540
 - internal agent reasoning, 540–543
 - performance criteria, 532–533
 - toploader, 534–538
 - worst-case scenario, 533–534
 - equipment for, 504–506
 - automated guided vehicles, 505
 - autonomous vehicle storage and retrieval systems, 505
 - choosing, 506
 - conveyors, 504
 - hoists, cranes, jibs, 505
 - palletizers, 504
 - robots, 505
 - trucks, 504
 - warehouse MHDs, 505
 - intelligent control of, *see* Intelligent control of MHSs
 - multiobjective model for operation allocation/selection in FMS design, 506–508
- principles of, 498–504
 - automation, 502–503
 - environment, 503
 - ergonomics, 499
 - life cycle, 503
 - planning, 498
 - space utilization, 500–501
 - standardization, 498
 - system, 501–502
 - unit load, 500
 - work, 499
- warehousing, 509–510

- Materials requirements planning (MRP), 77–86
 - defined, 77
 - definitions related to, 77–78
 - lot sizing techniques, 85–86
 - procedures and required inputs, 78–85
- Material safety data sheets (MSDS), 723
- Mathematical modeling (safety), 701
- Matrix (agents), 531
- Matrix diagram, 631, 633
- Matrix structures, 7, 8
- Mattel, 582
- Maturity, organizational, 12
- Maximal use of resources principle (TRIZ), 367
- Maximum lateness/maximum tardiness (single-machine problem), 89
- Maximum shortage (inventory), 64
- MCM (management coefficient models), 74, 75
- MDA (U.S. Missile Defense Agency), 674–675
- Mean flow time (single-machine problem), 88
- Mean lateness (single-machine problem), 89
- Measurement, 10–12, 14–15
 - and product/customer needs, 12
 - and quality control, 325
 - system implementation results, 23–24
 - TRIZ measurement and detection standards, 379–383
- Measurement systems analysis (MSA), 639–640
- Measure phase (DMADV), 657–660
- Measure phase (DMAIIC), 639–643
- Mechanical face seals, 296, 298–301
 - computer analysis tools for, 305, 308
 - emissions performance of, 301
 - leakage, 299–301
 - materials selection for, 300–301
 - seal balance, 296, 298–299
 - seal face flatness, 300
- Mechanical fasteners, 255–282
 - assembly features/functions, 256–259
 - bolted and riveted joints:
 - clamping force upper limits, 271–272
 - complex butt joint (bearing-type connection), 266–268
 - efficiency, 264
 - friction-type connections, 268–270
 - simple lap joint strength, 264–265
 - slip characteristics evaluation, 276–277
 - theoretical behavior of joint under tensile loads, 272–276
 - types, 262–264
 - bolts and rivets, 261–262
 - bolt tightening, 278–280
 - design for cyclical tension loads, 280–282
 - fatigue failure, 280–282
 - nesting strategy, 259–260
 - three-part assemblies, 260–261
- Medical devices standard (ISO 13485), 673–674
- Medium-frequency (MF) ac magnetron sputter deposition, 243–244
- MEDRAD, 636
- MEMS (microelectromechanical systems), 445
- Mendeley, 819
- METADEX, 815
- Metal casting and molding, 218–224
 - centrifugal casting, 220–221
 - environmentally benign manufacturing, 32–38
 - die casting, 37–38
 - sand casting, 32–36
 - investment casting, 224
 - permanent-mold casting, 222–223
 - plaster mold casting, 223–224
 - sand casting, 218–220
- Metal-cutting processes:
 - economics of, 123, 125–126
 - environmentally benign, 31
 - principles, 116–119
- Metal forming and shaping, 195–217
 - cold-working processes, 207–217
 - bending, 208, 210–213
 - classification of, 208
 - drawing, 208, 215–217
 - shearing, 208, 213–215
 - squeezing, 208–210
 - environmentally benign manufacturing, 38–41
 - hot-working processes, 195–207
 - classification of, 197
 - drawing, 203–205
 - extrusion, 201–202
 - forging, 199–201
 - piercing, 207
 - pipe welding, 206–207
 - rolling, 197–199
 - spinning, 206
 - powder metallurgy, 226–227
- Metal-joining processes:
 - brazing and soldering, 45
 - environmentally benign manufacturing, 41–45
 - welding, 41–44
- Metallic gaskets, 285, 287
- Metallizing, 231
- Metals Handbook*, 825
- Metasearching, 807
- Metasystem level (manufacturing systems), 185–187
- Metric Standards for Engineers*, 826
- MF (medium-frequency) ac magnetron sputter deposition, 243–244
- MHDs, *see* Material handling devices

- MHSs, *see* Material handling systems
 Microelectromechanical systems (MEMS), 445
 Microlevel (process planning), 188
 Milling machines, 140, 143
 Milling processes, 122, 140–143
 MIL-Q-9858 (Quality Program Requirements), 668
 Minimal attractive rate of return (MARR), 591–592, 601, 609
 Misdesignation of inventorship, 780
 Misjoinder of inventors, 780
 Misrepresentation, 759
 Mission assurance, 674–675
 Mission assurance provisions (MAPs), 668, 674–675
 Mitsubishi Heavy Industries, 657
 Mixed-model assembly line, 101–103
 MMPC (multipolar magnetic plasma confinement), 247
 Model-based definition, STEP for, 392, 394
 Model-based manufacturing, STEP for, 394–395
 Modeling (DMADV), 661
 Molding:
 metals, *see* Metal casting and molding
 plastics:
 injection molding processes, 45–50
 molding processes, 224–226
 Money:
 equivalence of transactions, 584–588
 as motivator, 570
 time value of, 583–584
 Monitoring (DMAIC), 653
 MOOC (massive open online course), 808, 820
 MORT (management oversight and risk tree), 700
 Motivation, 564–569
 as function of risks and challenges, 566–567
 implications for engineering performance, 566
 managing in high-motivation range, 567–569
 understanding motivational needs, 574
 Moving average, 55–56
 Moving magnets, 247
 MRP, *see* Materials requirements planning
 MRP-II (manufacturing resources planning), 77
 MRP schedule, 78–80
 MSA (measurement systems analysis), 639–640
 MSDS (material safety data sheets), 723
 Muda (waste), 619–620
 Mulally, Alan, 7
 Multiagent-based systems (MASs), 520–523. *See also* Agents
 Multicavity die, 37
 Multigenerational planning (DMADV), 656
 Multiobjective model (FMS), 506–508
 Multiple acceptance sampling, 336
 Multipolar magnetic plasma confinement (MMPC), 247
 Multiunit processes, 402–403
- ## N
- Nanohub, 816
 Nanotechnologies, 18–19
 NASA, 674–675, 677
 National Emission Standards for Hazardous Air Pollutants (NESHAP), 301
 National Environmental Policy Act (NEPA, 1969), 694
 National Housing Quality (NHQ) Award, 686
 National Institutes of Health (NIH), 820
 National Institute of Standards and Technology (NIST), 680, 684, 685
 National quality standards, 669. *See also* Quality management systems
 National Science Foundation (NSF), 820
 National Society of Professional Engineers (NSPE) code of ethics, 582–583
 NC, *see* Numerical control
 NDE (nondestructive evaluation), 441
 NDI, *see* Nondestructive inspection
 Need analysis (PPE), 725
 Negligence:
 and assumption of risk defense, 768
 and breach of contract, 756
 contributory/comparative, 766
 product liability, 758
 for services, 755
 and standard of care, 751–752
 Negotiation (agent communication), 527
 NEPA (National Environmental Policy Act, 1969), 694
 NESHAP (National Emission Standards for Hazardous Air Pollutants), 301
 Nesting strategy (mechanical fasteners), 259–260
 Net present value (NPV), 608–609
 Net requirements, 78, 81
 Neutron radiography, 450, 451
New America Icon (Bryce Hoffman), 7
 NFPA, 720
 NFPA “hazard diamond,” 720–722
 NHQ (National Housing Quality) Award, 686
 Nibbling (metals), 215
 NIH (National Institutes of Health), 820
 NIST, *see* National Institute of Standards and Technology
 Noise hazard, 713
 Noise variables, 399
 Nominal group techniques, 55

- Nominal interest rate, 585
 - Noncompetition provisions, 754
 - Noncontacting seals (high-speed/aerospace applications), 304–308
 - Nondestructive evaluation (NDE), 441
 - Nondestructive inspection (NDI), 441–476
 - adhesive bond NDE case study, 474–476
 - commonly-used methods, 444
 - eddy current inspection, 469–475
 - impedance plane, 470–473
 - lift-off of inspection coil from specimen, 472–474
 - probes and sensors, 474
 - skin effect, 470
 - future capabilities, 443, 445
 - information sources on inspection methods, 442–443
 - instrumentation qualities, 442
 - liquid penetrants, 445–447
 - magnetic particle inspection, 465–468
 - radiography, 448–456
 - attenuation of X-radiation, 450–452
 - computed tomography, 455–456
 - film-based radiography, 452–453
 - generation/absorption of X-rays, 448–450
 - neutron radiography, 450, 451
 - penetrameter, 453–454
 - real-time radiography, 454–455
 - thermal inspection, 468–469
 - ultrasonic inspection, 456–465
 - bond testers, 464–465
 - inspection process, 461–464
 - reflection and transmission of sound, 458–459
 - refraction of sound, 459–461
 - sound waves, 457–458
 - ultrasonic properties of common materials, 476–493
 - Nonmetallic bristles (brush seals), 318
 - Nonmetallic gaskets, 285
 - Nonobviousness (patentability), 782–783
 - Nontraditional machining, 158–180
 - abrasive flow machining, 158–159, 162
 - abrasive jet machining, 159, 162
 - chemical machining (chemical milling, chemical blanking), 178
 - electrical discharge grinding, 173
 - electrical discharge machining, 173, 174
 - electrical discharge sawing, 174
 - electrical discharge wire cutting (traveling wire), 174–175
 - electrochemical deburring, 166
 - electrochemical discharge grinding, 166, 167
 - electrochemical grinding, 167
 - electrochemical honing, 167–168
 - electrochemical machining, 168–169
 - electrochemical polishing, 169, 170
 - electrochemical sharpening, 169, 170
 - electrochemical turning, 170, 171
 - electromechanical machining, 160
 - electron beam machining, 172–173
 - electropolishing, 178, 179
 - electrostream, 170, 171
 - hydrodynamic machining, 159, 163
 - laser beam machining, 175–176
 - laser beam torch, 176–177
 - low-stress grinding, 159, 160, 163
 - photochemical machining, 178–179
 - plasma beam machining, 177, 178
 - rapid prototyping/rapid tooling, 180
 - shaped-tube electrolytic machining, 171–172
 - thermally assisted machining, 160, 164
 - thermochemical machining, 179–180
 - total form machining, 161, 164
 - ultrasonic machining, 161, 163, 165
 - water-jet machining, 163–165
 - Nose radiuses, 127
 - Notching, 213, 215
 - Novelty (patentability), 780–782
 - np* charts, 333–334
 - NPV (net present value), 608–609
 - NSF (National Science Foundation), 820
 - NSPE (National Society of Professional Engineers) code of ethics, 582–583
 - NTIS, 815
 - Numerical control (NC), 344–348
 - adaptive control, 347, 348
 - CAD/CAM part programming, 346–347
 - components, 344–345
 - coordinate system, 345–346
 - machinability data prediction, 347
 - numerical control components, 344–345
 - parts selection for machining, 346
 - programming by scanning and digitizing, 347
 - STEP-NC³, 394
- O**
- Objectives, communicating, 576
 - Oblique cutting, 116, 118
 - Occupational Safety and Health Act (OSHAct, 1970), 692, 697, 698, 739
 - Occupational Safety and Health Administration (OSHA), 697–700
 - legal impact of standards, 755
 - MSDSs requirements, 723
 - occupational safety and health standards, 697–698, 744

- Occupational Safety and Health Administration
 - (*continued*)
 - required notices and records, 698
 - safety design requirements, 723–724
 - state-operated compliance programs, 698–700
 - OFAAT (one-factor-at-a-time), 645–646
 - Ohno, Taiichi, 618, 619
 - Oil Pollution Act (OPA, 1990), 696
 - Oldsmobile Motor, 516
 - One-factor-at-a-time (OFAAT), 645–646
 - Online (term), 808
 - Online information resources, 805–821
 - access options, 809–811
 - database services, 813–817
 - document delivery options, 817–818
 - future of online access, 819–821
 - indexes and abstracts, 825
 - managing online literature, 819
 - for nondestructive inspection, 443
 - options for using, 821
 - security issues, 811–812
 - terms related to, 806–809
 - OPA (Oil Pollution Act, 1990), 696
 - Open-die hammer forging (metals), 200
 - Open-source software, 807, 808
 - Open structures, 519
 - OpenURL, 808
 - Operating system (OS), 808, 812
 - Operation characteristic curves (control charts), 332
 - Oral contracts, 755
 - Ordering cost, 64
 - Organic coatings, 231
 - Organizational culture, 563, 577
 - Organizational interfaces, 575
 - Organizational maturity, 12
 - Organization structures:
 - adaptability of, 14
 - and empowerment of workforce, 16
 - hierarchical, 7
 - ideal, 8
 - for manufacturing, 10–12
 - open and dynamic, 519
 - and product/customer needs, 12
 - and workforce reductions, 16–17
 - Originality of inventorship (patentability), 779–780
 - O-rings, 289–292
 - basic sealing mechanism, 289–290
 - material selection/chemical compatibility, 291
 - preload, 290–291
 - rotary applications, 291–292
 - thermal effects, 291
 - Orthogonal cutting, 116, 118
 - OS (operating system), 808, 812
 - OSHA, *see* Occupational Safety and Health Administration
 - OSHAct, *see* Occupational Safety and Health Act, 1970
 - Output variables, 399–400
 - Overlapping stress concentrations (mechanical fasteners), 281
 - Oxide tool inserts, 127
- P**
- PAC (plasma-arc cutting), 178
 - Pacifico project rating factor, 610
 - Packings and braided rope seals, 292–295
 - PACO (coordinated patterns), 545–550
 - Paints, 231, 469
 - Palletizers, 504
 - PAPVD (plasma-assisted PVD), 236–238, 247
 - Parallel line balancing, 103–104
 - Pareto, Vilfredo, 644
 - Pareto charts, 627, 629, 644
 - Pareto plots, 15
 - Pareto principle, 644
 - Paris Convention, 702
 - Parkerizing, 233
 - Parts classification and coding (group technology), 354–357
 - Part family formation (group technology), 354
 - Patents, 754, 773–803
 - application types, 775
 - enforcing against infringers, 797–801
 - alternative resolution of disputes, 800
 - declaratory judgment actions, 799
 - defenses to enforcement, 798
 - failure to sue infringers, 799
 - government infringement, 800
 - interferences, 800–801
 - outcome of suits, 798–799
 - settling suits, 799
 - types of infringement, 797–798
 - issuance of, 795–796
 - “negative right” granted by, 776–777
 - preparing to apply for, 783–789
 - application preparation, 785
 - best mode requirement, 786
 - claims format, 786–787
 - description requirement, 786
 - distinctness requirement, 786
 - enablement requirement, 785–786
 - executing applications, 787–788
 - express mail filing, 789

- patentability search, 784
- product-by-process claims, 786
- putting invention in proper perspective, 784
- small-entity status qualification, 788–789
- U.S. Patent and Trademark Office fees, 788
- prosecution of pending applications, 789–797
 - continuation, divisional, and continuation-in-part applications, 794–795
 - double-patenting rejections, 794
 - duty of candor, 791
 - initial review of application, 791–792
 - interviewing examiner, 793–794
 - maintaining chain of pending applications, 795
 - “Patent Pending” indication, 789–790
 - publishing pending applications, 790
 - reconsideration in view of filing of response, 793
 - response to office actions, 792–793
 - restriction and election requirements, 794
- protections outside of the United States, 801–803
- provisional, 775–776
- reexamination, 796–797
- reissue to correct errors, 796
- requirements for patentability, 777–783
- safeguarding original patent document, 796
- terms and expiration, 774–775
- types of, 774
- Patentability requirements, 777–783
 - ideas and inventions vs. patentable inventions, 777–778
 - nonobviousness, 782–783
 - novelty, 780–782
 - originality of inventorship, 779–780
 - statutory bar requirements, 781, 783
 - statutory subject matter, 778–779
 - utility, 782
- Patentability search, 784
- Patent Cooperation Treaty (PCT), 802–803
- “Patent Pending” indication, 789–790
- Payback period, 610
- PBM (plasma beam machining), 177, 178
- PBMA (Process Based Mission Assurance) plan, 674
- p* charts, 332–334
- PCM (photochemical machining), 178–179
- PCT (Patent Cooperation Treaty), 802–803
- PDCA cycle, *see* Plan–do–check–act cycle
- pdf (Portable Document Format), 808
- PDPC (process decision program chart), 631, 633
- PDSA (plan–do–study–act), 621
- Peening (metals), 210
- Pending patents, 789–797
 - continuation, divisional, and continuation-in-part applications, 794–795
 - double-patenting rejections, 794
 - duty of candor, 791
 - initial review of application, 791–792
 - interviewing examiner, 793–794
 - maintaining chain of pending applications, 795
 - “Patent Pending” indication, 789–790
 - publishing pending applications, 790
 - reconsideration in view of filing of response, 793
 - response to office actions, 792–793
 - restriction and election requirements, 794
- Penetrameter, 453–454
- Penetrant testing, *see* Liquid penetrants (NDI)
- People, managing, *see* Management, of people
- Perforating (metals), 215
- Performance, motivation and, 566
- Performance characteristics, 400
- Performance excellence awards, *see* Quality and performance excellence awards
- Periodicals, 824
- Periodic interest rate, 585
- Periodic order quantity (POQ), 85–86
- Permanent-mold casting (metals), 37, 222–223
- Perpendicularity (mechanical fasteners), 281
- Personal protective equipment (PPE), 724–726
 - and engineering controls, 715
 - training for use of, 742
- Peters, Tom, 624
- Petroleum solvents (cleaning), 228–229
- PFA (production flow analysis), 357
- Phosphate coatings, 233
- Photochemical machining (PCM), 178–179
- Physical contradictions (TRIZ), 366–367, 371, 373
- Physical protection training, 742
- Physical vapor deposition (PVD), 235–249
 - chemical vapor deposition vs., 235
 - film formation and growth, 237, 239
 - glow discharge plasma, 236–238
 - processes, 239–249
 - beam, 247–249
 - evaporative, 239–242
 - sputter deposition, 242–247
- Pickling (cleaning), 229
- Piercing (metals), 207, 213, 215
- Piloting the solution, 650
- Pipe welding, 206–207
- Pitch (joints), 262
- Pitch-catch inspection, 462–464
- Plaintiff (lawsuits), 751
- Plan (term), 498

- Plan–do–check–act (PDCA) cycle, 186, 187, 620–621
- Plan–do–study–act (PDSA), 621
- Planing, 149, 152
- Planned order receipts, 81
- Planned order releases, 81, 85
- Planning:
 - in engineering management, 575
 - of manufacturing systems assessments, 189–190
- Planning principle (material handling), 498
- Plant application (patents), 775
- Plant patents, 774, 775
- Plasma-arc cutting (PAC), 178
- Plasma arc welding, 43
- Plasma-assisted PVD (PAPVD), 236–238, 247
- Plasma beam machining (PBM), 177, 178
- Plaster mold casting (metals), 223–224
- Plastics:
 - as coatings, 230
 - machining, 153
- Plastic deformation (bolts), 270
- Plastic injection molding, 45–50
 - auxiliaries, 47, 224
 - environmental analysis of, 48–50
 - equipment, 46
 - life cycle of products, 49–50
 - materials, 47
 - principle of, 45–46
 - products, 47–48
 - tooling, 46–47
- Plastic molding processes, 224–226
 - blow molding, 225–226
 - coinjection molding, 225
 - expandable-bead molding, 225
 - extruding, 225
 - forged-plastic parts, 226
 - injection molding, 45–50, 224
 - reinforced-plastic molding, 226
 - rotomolding, 225
 - thermoforming, 226
- Plating:
 - electroplating, 232
 - hot-dip plating, 232
 - ion plating, 237
 - plastic injection molding, 47
- PLCS (Product Life Cycle Systems), 396
- PM (powder metallurgy), 226–227
- Point-to-point machine tool (NC), 345
- Point-to-point robot systems, 349
- Poka-yoke, 623
- Polishing, 230
 - electrochemical, 169, 170
 - electropolishing, 178, 179
- Pollution Protection Act (PPA, 1990), 697
- Polycrystalline diamond tools, 127
- POQ (periodic order quantity), 85–86
- Portable Document Format (pdf), 808
- Powder metallurgy (PM), 226–227
- Power profile (in management), 569–570
- Power spinning, 206
- PPA (Pollution Protection Act, 1990), 697
- PPE, *see* Personal protective equipment
- Precision-casting processes, 37
- Predictor (TRA), 412, 414
- Preload:
 - bolts, 269
 - O-rings, 290
- Present value (PV), 590–592
- Present worth (PW), 609
- Press forging (metals), 200
- Pressure closing (brush seals), 316
- PR (project rating) factor, 610
- Primary literature, 823–824
- The Principles of Scientific Management* (Frederick Taylor), 6
- Prioritization matrix, 629, 631, 632
- Priority dispatching rules (job sequencing/scheduling), 94–97
- Probabilistic inventory models, 65
- Process(es):
 - as legal patent term, 778
 - in manufacturing systems, 12
 - production, *see* Production processes and equipment
- Process Based Mission Assurance (PBMA) plan, 674
- Process capability analysis, 640–643
- Process capability studies, 406
- Process decision program chart (PDPC), 631, 633
- Process Excellence, 636. *See also* Total quality management (TQM)
- Process layout (production plants), 342
- Process planning, 188, 358
- Process selection (environmentally conscious manufacturing), 188
- Process simulation (DMADV), 662–663
- Process technology, 397–433
 - definitions related to, 435–436
 - history of technological development, 397–398
 - nontechnical statistical glossary, 436–440
 - traditional approach, 398–406
 - inefficiencies with, 406
 - problems with, 401–406
 - Tuszynski's process law, 433–435

- Tuszynski's relational algorithm
 - (new technology), 406–432
 - across engineering functionalities, 408
 - correlation charts, 412, 415, 416
 - fixes, 424–428
 - foundation of, 409
 - generating input data for, 411–413
 - graphical illustration of, 409, 410
 - history of, 406–407
 - material selection, 428–432
 - operating point adjustments, 411
 - process conclusions for, 409, 410
 - region of conformance, 414, 416
 - relationship conditions in, 414, 416–425
 - selecting predictors, 412, 414, 415
 - as single-degree-of-freedom system, 410–411
 - usefulness of, 409
 - what it is not, 407
- Procurement quantity, 64
- Producers, 341
- Producibility window, 404–405
- Products, 14
 - and environmentally benign manufacturing, 50
 - extended, 19
 - life cycle, 49–50
 - STEP for, 392, 393
 - sustainability and, 18
 - responsibility for, through life cycle, 510
- Product-by-process patent claims, 786
- Product defects:
 - nature of, 760–762
 - design flaws, 760–761
 - instructions and warnings, 761–762, 764
 - production or manufacturing flaws, 760
 - uncovering, 762–764
 - design hierarchy, 763–764
 - hazard analysis, 762–763
 - hazard index, 763
- Product design and engineering function, 341
- Product development figure of merit, 610
- Product flow layout (production plants), 342
- Production (in supply chain management), 108
- Production 2000+ (P2000+), 518
- Production costs, 74
- Production flaws, 760
- Production flow analysis (PFA), 357
- Production kanban, 105
- Production operations, 341–342
- Production operation models, 342–344
- Production planning, 53–110
 - aggregate planning, 73–77
 - approaches to, 74–75
 - costs of, 74
 - lack of adoption of, 76–77
 - levels of aggregation/disaggregation, 76
 - meeting demand fluctuations, 74
 - forecasting, 54–63
 - causal methods, 58–60
 - definitions related to, 54–55
 - error analysis, 61, 63
 - qualitative, 55–58
 - time series analysis, 60–62
 - inventory models, 63–73
 - definitions related to, 65–66
 - modeling approach, 66–73
 - types of, 65
 - Japanese manufacturing philosophy, 104–107
 - job sequencing and scheduling, 87–104
 - assembly line balancing, 97–104
 - flow shops, 90–92
 - heuristics/priority dispatching rules, 94–97
 - job shops, 92–94
 - single-machine problem, 88–90
 - structure of sequencing problem, 87–88
 - materials requirements planning, 77–86
 - definitions related to, 77–78
 - lot sizing techniques, 85–86
 - procedures and required inputs, 78–85
 - supply chain management, 108–110
- Production planning and control function, 342
- Production plants/facilities, 342
- Production processes and equipment, 115–180
 - broaching, 148–151
 - cutting-tool materials, 126–129
 - drilling machines, 133–140
 - gear manufacturing, 143–146
 - grinding, abrasive machining, and finishing, 153–158
 - abrasives, 154–156
 - temperature, 156–158
 - machining plastics, 153
 - machining power and cutting forces, 119–121
 - metal-cutting:
 - economics in, 123, 125–126
 - principles, 116–119
 - milling processes, 140–143
 - nontraditional machining, 158–180
 - abrasive flow machining, 158–159, 162
 - abrasive jet machining, 159, 162
 - chemical machining (chemical milling, chemical blanking), 178
 - electrical discharge grinding, 173
 - electrical discharge machining, 173, 174
 - electrical discharge sawing, 174
 - electrical discharge wire cutting (traveling wire), 174–175
 - electrochemical deburring, 166

- Production processes and equipment (*continued*)
 - electrochemical discharge grinding, 166, 167
 - electrochemical grinding, 167
 - electrochemical honing, 167–168
 - electrochemical machining, 168–169
 - electrochemical polishing, 169, 170
 - electrochemical sharpening, 169, 170
 - electrochemical turning, 170, 171
 - electromechanical machining, 160
 - electron beam machining, 172–173
 - electropolishing, 178, 179
 - electrostream, 170, 171
 - hydrodynamic machining, 159, 163
 - laser beam machining, 175–176
 - laser beam torch, 176–177
 - low-stress grinding, 159, 160, 163
 - photochemical machining, 178–179
 - plasma beam machining, 177, 178
 - rapid prototyping/rapid tooling, 180
 - shaped-tube electrolytic machining, 171–172
 - thermally assisted machining, 160, 164
 - thermochemical machining, 179–180
 - total form machining, 161, 164
 - ultrasonic machining, 161, 163, 165
 - water-jet machining, 163–165
 - sawing, shearing, and cutting off, 152–153
 - shaping, planing, and slotting, 149, 152
 - thread cutting and forming, 146–148
 - tool life, 121–124
 - turning machines, 129–133
 - Production rate change costs, 74
 - Production scheduling, 188
 - Production switching heuristics (PSH), 74, 75
 - Production systems, *see* Manufacturing systems
 - Productivity, 8–10, 16
 - Product liability, 757–770
 - after-market hazards, 768–770
 - defense to, 764–768
 - laws of, 757–759
 - nature of product defects, 760–762
 - recalls, retrofits, and continuing duty to warn, 768–770
 - uncovering product defects, 762–764
 - Product life cycle, 18, 49–50, 392, 393
 - Product Life Cycle Systems (PLCS), 396
 - Product manufacturing information (PMI)), 392, 394
 - Product structure tree, 78
 - Professional interests, accommodating, 575
 - Professional liability, 751–757
 - business liability, 755–757
 - case study, 757
 - employee liability, 751–754
 - joint and severable, 752
 - Project charter, 637–638, 656
 - Project evaluation and selection, 605–616
 - management perspective on, 605–607
 - qualitative approaches, 611–612
 - quantitative approaches, 607–611
 - cost–benefit, 610
 - net present value, 608–609
 - payback period, 610
 - product development figure of merit, 610
 - project rating factor, 610
 - return on investment, 609
 - recommendations for, 613–615
 - terms related to, 615
 - variables and abbreviations, 616
 - Project formulation, in assessment of systems, 192–193
 - Project plan, 656
 - Project rating (PR) factor, 610
 - Project teams, 571–573
 - Prototyping, 180
 - Provisional application (patents), 775–776
 - Provisional rights (published patents), 790
 - PSH (production switching heuristics), 74, 75
 - Publication of pending patent applications, 790
 - Pulsed laser deposition, 249
 - Pulsed power sputtering (PVD), 243–244
 - Pulse-echo inspection, 462–464
 - Pulse-echo instruments (bolts), 279–280
 - Punishment, as basis of interpersonal power, 569
 - Purchase discounts, 69–73
 - Purchase model with shortage permitted (inventory), 68
 - Purchase model with shortage prohibited (inventory), 66–67
 - Purchasing (in supply chain management), 108
 - Purchasing power, inflation and, 600–601
 - PV (present value), 590–592
 - PVD, *see* Physical vapor deposition
 - PW (company), 319
 - PW (present worth), 609
- ## Q
- QFD, *see* Quality function deployment
 - QMS (Quality Management System) registrars, 670
 - QS 9000 (automotive industry), 673
 - Quadratic regression, 60
 - Qualitative forecast (term), 55
 - Qualitative forecasting, 55–58
 - exponential smoothing, 57–58
 - moving average, 55–56
 - weighted moving average, 56–57

- Quality and performance excellence awards, 678–688
 - Baldrige National Quality Award, 680–684, 687
 - comparison of, 686–688
 - Deming Prize, 686, 687
 - industry-specific quality awards, 686
 - programs around the world, 685–686
 - Shingo Prize for Operational Excellence, 684–685
 - U.S. state quality awards, 684, 687
 - Quality control, 326. *See also* Quality management systems
 - measurements and, 325
 - statistical, *see* Statistical quality control
 - traditional tools for, 624–629
 - Quality control function, 342
 - Quality function deployment (QFD), 15, 639, 652, 657–660
 - Quality loss function, 649
 - Quality management systems, 667–688.
 - See also* Total quality management (TQM)
 - accreditation, 668–669
 - Capability Maturity Model Integration, 675, 676
 - certification, 668–670
 - mission assurance, 674–675
 - national and international standards, 669
 - quality and performance excellence awards, 678–688
 - award programs around the world, 685–686
 - Baldrige National Quality Award, 680–684, 687
 - comparison of, 686–688
 - Deming Prize, 686, 687
 - industry-specific quality awards, 686
 - Shingo Prize for Operational Excellence, 684–685
 - U.S. state quality awards, 684, 687
 - registration, 668, 670, 671
 - standards:
 - AS 9100: aviation, space, and defense organizations, 670, 672, 673
 - ISO 9001, 669–672
 - ISO 13485: medical devices, 673–674
 - ISO 14000: environmental management systems, 675–677
 - ISO 22000: food safety management, 677–678
 - ISO/TS 16949: automotive production/service part organizations, 672, 673
 - TL 9000: telecom, 674
 - Quality Management System (QMS) registrars, 670
 - Quantitative forecast, 55
 - Quantity discount with variable holding cost, 71–73
- ## R
- Radial lip seals, 296
 - Radio-frequency identification (RFID), 445
 - Radiography testing (RT), 448–456
 - attenuation of X-radiation, 450–452
 - capabilities, 444
 - computed tomography, 455–456
 - film-based radiography, 452–453
 - generation/absorption of X-rays, 448–450
 - neutron radiography, 450, 451
 - penetrameter, 453–454
 - real-time radiography, 454–455
 - Rake angles, 127, 213
 - Random experimentation, 647
 - Randomization (experiments), 649
 - Rapid prototyping/rapid tooling, 180
 - Rattling (abrasive barrel finishing), 229
 - R chart, 326–331
 - RCRA (Resource Conservation and Recovery Act, 1976), 694–695
 - Reactive agents, 523
 - Real interest rate, 601
 - Real-time radiography (RTR), 454–455
 - Reasonable foreseeable misuse (product liability), 762, 766
 - Reasonableness test (negligence), 758
 - Recalls, 768–770
 - Reciprocating applications (O-rings), 291–292
 - Reciprocating power hacksaw, 152
 - Reductions in workforce, 16–17
 - Reengineering, 16
 - Reexamination of patent validity, 796–797
 - Referent power, as basis of interpersonal power, 569
 - Reflection (sound), 458–459
 - Refraction (sound), 459–461
 - Registration (quality management systems), 668, 670, 671
 - Regression analysis, 649
 - basic, 58–59
 - defined, 55
 - quadratic regression, 60
 - simple linear regression, 59
 - Reinforced-plastic molding, 226
 - Reissue patents, 796
 - Reliability assessment (DMADV), 661
 - Relief angles, 127
 - Remote login, 808

Reorder point, 65
 Reorder quantity, 65
 Repeating sections (joints), 263
 Replacement studies, 599
 Replenishment rate, 66
 Replication (experiments), 649
 Reporting, in assessment of systems, 189, 190, 192–193
 Resistance-welding processes, 41, 43
 Resources:
 effective use of, 18–19
 information, *see* Information sources/resources
 maximal use of, 367
 in TRIZ, 367
 waste of, 184
 Resource Conservation and Recovery Act (RCRA), 1976), 694–695
 Restoration, 19
 Restriction requirement (patents), 794
 Retrofits, 768–770
 Return on investment (ROI), 609
 Reverse supply chain (reverse logistics) management, 510
 Reward:
 as basis of interpersonal power, 569
 creating systems of, 576
 rf diode sputtering (PVD), 243
 RFID (radio-frequency identification), 445
 Risk:
 assumption of, 767–768
 motivation as function of, 566–567
 Risk analysis, 15, 602
 Risk assessment, 707–708
 Risk management, 650, 661
 Risk register, 657
 Risk score, 707–708
 Rivets, 261–262. *See also* Mechanical fasteners
 Riveting (metals), 210
 Robots:
 in assembly, 516, 517
 components, 349
 industrial, 349–351
 material handling, 505
 programming systems, 350
 Robust design, 649
 ROI (return on investment), 609
 Roll bending (metals), 212
 Rolled threads (bolts), 280–281
 Roller leveling, 213
 Roll forging (metals), 201
 Rolling:
 abrasive barrel finishing, 229
 threads, 146, 148
 Rolling (metals), 197–199

Roll straightening, 213
 Rope packings, 292–295
Roper Corp. v. Litton Systems, Inc., 796
 Rotary applications (O-rings), 291–292
 Rotary ultrasonic machining (RUM), 159
 Rotatable magnetron, 247
 Rotomolding (plastics), 225
 Rotordynamic stability, labyrinth seals and, 311
 Royalties, 787
 RT, *see* Radiography testing
 RTR (real-time radiography), 454–455
 Rubber bag forming, 217
 RUM (rotary ultrasonic machining), 159
 Runner molding, 47
 Rupture stress (bolts), 270

S

SAE TechSelect, 818
 Safety engineering, 691–745
 alternatives to engineering controls, 715, 717–719
 design and redesign, 719–724
 hardware, 719, 720
 hazardous material classification system, 720–722
 material safety data sheets, 723
 process, 719, 720
 safety design requirements, 723–724
 employee needs and expectations, 692–693
 engineering controls for machine tools, 710–713
 government regulatory requirements, 693–700
 Environmental Protection Agency, 694–697
 Occupational Safety and Health Administration, 697–698
 state-operated compliance programs, 698–700
 human factors engineering/ergonomics, 708–711
 machine safeguarding methods, 714–717
 managing safety function, 727–735
 accident prevention, 727–729
 eliminating unsafe conditions, 729–734
 management principles, 728, 730
 supervisor's role, 727
 unsafe conditions checklist for mechanical or physical facilities, 734–735
 personal protective equipment, 724–726
 safety factor in, 750–751
 safety training:
 automated external defibrillator, 735
 driver, 735–736

- environmental risk, 736–739
- fire protection, 739
- first-aid, 739
- hazard recognition, 739–740
- HAZWOPER, 740
- job hazard analysis, 742–744
- lockout box, 740–741
- machine guard, 741
- management oversight of, 744–745
- personal protection equipment/physical protection, 742
- sources/types of training materials, 745
- specialized courses, 735–742
- system safety, 700–708
 - analysis methods, 700–702
 - fault tree technique, 702
 - preparation/review of procedures criteria, 702–707
 - risk assessment process, 707–708
- Safety factor, 750–751
- Safety standards:
 - as defense to product liability, 765–766
 - Hazard Communication Standard, 723
 - HAZWOPER, 740
 - legal impact of, 755
 - OSHA, 697–698, 744
- Safety training, 735–745
 - automated external defibrillator, 735
 - driver, 735–736
 - environmental risk, 736–739
 - fire protection, 739
 - first-aid, 739
 - hazard recognition, 739–740
 - HAZWOPER, 740
 - job hazard analysis, 742–744
 - lockout box, 740–741
 - machine guard, 741
 - management oversight of, 744–745
 - personal protection equipment/physical protection, 742
 - sources/types of training materials, 745
 - specialized courses, 735–742
- Salary, as motivator, 570
- Sales and marketing function, 341
- Sand casting (metals), 32–36, 218–220
- Sanden International, Inc., 680
- Sanitation, 726
- SARA (Superfund Amendments and Reauthorization Act, 1986), 696
- Saturn Project (GM), 17
- Sawing, 152–153, 174
- Scalability, 519
- Scanning, programming by, 347
- Scatter diagrams, 628, 629
- Scheduled receipts, 78, 81
- Scheduling, 87, 188, 517–518. *See also* Job sequencing and scheduling
- Schumpeter, Joseph, 8
- SCM (supply chain management), 86, 108–110
- Scopus, 816
- Seal balancing, 296, 298–299
- Seal hysteresis (brush seals), 315, 316
- Seal technology, 283–319
 - dynamic seals, 296–319
 - brush seals, 313–319
 - for emissions control, 301–305
 - honeycomb seals, 312–313
 - initial selection of, 296–298
 - labyrinth seals, 308–312
 - mechanical face seals, 296, 298–301
 - noncontacting seals for high-speed/aerospace applications, 304–308
 - radial lip seals, 296
 - turbine engine seals, 296
 - ongoing developments with, 319
 - static seals, 283–295
 - gaskets, 283–289
 - O-rings, 289–292
 - packings and braided rope seals, 292–295
- Seaming (metals), 213
- Search techniques (ST), 74, 75
- Seasonal data, 55, 60–62
- Seasonal forecasts, 61
- Seasonal movements (time series), 60
- Secrecy order (patents), 779
- Security, online, 811–812
- SEI (Software Engineering Institute), 675
- Selecting projects, *see* Project evaluation and selection
- Self-acting face seals, 304, 305, 307
- Self-directed teams, 572–573
- Self-direction, promoting, 576
- Self-fulfillment prophesy, 566–567
- Self-managed teams, 8–9
- Semantic Web, 820
- Semiconductor-based inspection devices, 445
- Sensitivity analysis, 601–602
- Sequencing, 87. *See also* Job sequencing and scheduling
- Sequential acceptance sampling, 336
- Server, 808
- Setup cost, 64
- SFM, *see* Substance field modeling
- Shallow drawing, 216
- Shaped-tube electrolytic machining (STEM), 171–172
- Shaping, 149, 152. *See also* Metal forming and shaping

- Sharing dependencies, 526
- Sharpening, electrochemical, 169, 170
- Shaving, 213, 215
- Shearing (metals), 152–153, 208, 213–215
- Shear spinning, 206
- Shell drawing (metals), 216
- Shewhart, Walter, 621, 654, 655
- Shibboleth, 808, 821
- Shingo, Shigeo, 620, 684
- The Shingo House, 685
- Shingo Prize for Operational Excellence, 684–685
- Shipping and receiving function, 342
- Shipping and receiving inspections, 406
- Shop-floor models, 518
- Shortage, 66
- Shortage cost, 65
- Short pitch (joints), 262
- Side cutting edge angle, 127
- σ chart, 326–330
- Silicon carbide (abrasive), 154
- Simulation (DMADV), 661–663
- Simulation models (SM), 74, 75
- Sinclair, Upton, 6
- Single-crystal diamonds, 127
- Single-machine problem, 88–90
- SIPOC, 638, 657
- Site visits, 191
- Six Sigma, 15, 16, 621, 636. *See also* Total quality management (TQM)
- Sizing (metals), 209
- SL (straight-line) depreciation, 600
- Sleeping agents, 550–552
- Slip, 231
- Slip-resistant joints, 276–277
- Slotting, 149, 152
- SM (simulation models), 74, 75
- Small companies, 21–22
- Small-entity status (patents), 788–789
- Smith Adam, 6
- Smoothing:
 - defined, 55
 - exponential, 57–58
- Soaking (hot rolling), 197
- Sobelman product development figure of merit, 610
- Social engineering, 15–17
- Social networks, 20
- Societies, as agents, 530
- Society of Automotive Engineers (SAE)
 - TechSelect, 818
- Software Engineering Institute (SEI), 675
- Soldering, 41, 45
- Solid-state detectors (NDI), 443
- Sombart, Werner, 8
- The Soul of a New Machine* (T. Kidder), 9
- Sound:
 - reflection and transmission of, 458–459
 - refraction of, 459–461
- Sound waves, 457–458
- Space organizations:
 - AS 9100 quality standard, 670, 672, 673
 - mission assurance, 674–675
- Space utilization principle (material handling), 500–501
- Spade drills, 139
- SPAM, 809, 812
- SPC, *see* Statistical process control
- Special cause variation, 654, 655
- Special hazards (NFPA hazard diamond), 722
- Species election requirement (patents), 794
- Specifications:
 - as information source, 826
 - for quality, 668
 - for ultrasonic testing, 457
- Spinning (metals), 206
- Sputter deposition processes (PVD), 242–247
 - dc diode sputtering, 242–243
 - magnetron sputtering, 245–247
 - pulsed power sputtering, 243–244
 - rf diode sputtering, 243
 - triode sputtering, 244–245
- Spyware, 809, 812
- SQC, *see* Statistical quality control
- Squeezing (metals), 208–210
- ST (search techniques), 74, 75
- Stability, 8–13
- Stakeholder analysis, 638
- Staking (metals), 210
- Standards:
 - for agent-based manufacturing control, 518
 - emissions control, 301–302
 - as information source, 826–827
 - information sources/resources, 826–827
 - quality management:
 - AS9100: aviation, space, and defense organizations, 670, 672–675
 - ISO 9001, 669–672
 - ISO 13485: medical devices, 673–674
 - ISO 14000: environmental management systems, 675–677
 - ISO 22000: food safety management, 677–678
 - ISO/TS 16949: automotive production/service part organizations, 672, 673
 - TL 9000: telecom, 674
- safety:
 - as defense to product liability, 765–766
 - Hazard Communication Standard, 723

- HAZWOPER, 740
 - legal impact of, 755
 - occupational safety and health, 697–698, 744
- Standard CMMI® Appraisal Method for Process Improvement (SCAMPI), 675
- Standard Handbook of Engineering Calculations*, 826
- Standardization principle (material handling), 498
- Standard of care (employee liability), 751–752
- States (U.S.):
 - Clean Air Act enforcement by, 696
 - Clean Water State Revolving Fund, 694
 - OSHA compliance programs, 698–700
 - patent infringement by, 800
 - quality awards, 684, 687
- State of the art, as defense to product liability, 764–765
- Static applications (O-rings), 291–292
- Static seals, 283–295
 - gaskets, 283–289
 - O-rings, 289–292
 - packings and braided rope seals, 292–295
- Station time, 98
- Statistical process control (SPC):
 - DMAIC control phase, 653–656
 - limitations of current studies, 406
 - and TRA, 410
- Statistical quality control (SQC), 325–336, 655–656
 - acceptance sampling, 335–336
 - control charts, 326–331
 - dimension and tolerance, 325
 - interrelationship of tolerances of assembled products, 331
 - measurements and, 325
 - operation characteristic curves, 332
- Statistical tolerancing (DMADV), 662
- Statutory bars (patentability requirement), 781, 783
- Statutory subject matter (patentability), 778–779
- Steel friction disks, 152–153
- STEM (shaped-tube electrolytic machining), 171–172
- STEP (Standard for Product Model Data), 391–396
 - application protocols, 391–392
 - for building information management, 395
 - for model-based definition, 392, 394
 - for model-based manufacturing, 394–395
 - for product life cycle, 392, 393
- STEP Tools, Inc., 392
- Stiglitz, Joseph, 24
- STN database service, 815
- Stone and Webster Engineering (S&W), 757
- Storage and retrieval systems, 505, 510–511
- Straightening (metals), 213
- Straight-line (SL) depreciation, 600
- Strategic level (manufacturing systems), 186, 188
- Stratification, 628, 629
- Stress cracking (bolts), 271
- Stretcher leveling, 213
- Stretch forming (metals), 216
- Strict liability, 758–759, 768
- Substance field modeling (Su-Field, SFMs), 375–376, 378–379
- Substandard conditions, 757
- Substitution, as safety control, 717–718
- Subsumption agent architecture, 524
- Success, as function of attitude, 567
- Su-Field, *see* Substance field modeling
- Superfinishing, 158
- Superfund Act (CERCLA), 695–696
- Superfund Amendments and Reauthorization Act (SARA, 1986), 696
- Suppliers, 30, 509
- Supply chain:
 - environmentally benign manufacturing, 30–31
 - kanban applied to, 109–110
 - vertical integration, 20
- Supply chain management (SCM), 86, 108–110
- Surface conditions (mechanical fasteners), 282
- Surface engineering (PVD), 235–249
 - chemical vapor deposition vs., 235
 - film formation and growth, 237, 239
 - glow discharge plasma, 236–238
 - processes, 239–249
- Surface finishing, 157–158, 227
- Surface treatment, 227–233
 - chemical conversions, 232–233
 - cleaning, 227–230
 - coatings, 230–232
- Sustainability, 18–19
 - biomimicry, 18–19
 - conservation and restoration, 19
 - effective use of resources, 18–19
 - extending manufacturers' responsibility, 19
- Svoye Logistics, 510–511
- S&W (Stone and Webster Engineering), 757
- Swaging (metals), 201, 209
- Symantec, 812
- System (term), 501
- Systems design, 13–15
- Systems evolution, laws of, 374–375
- Systems integration (FMS), 517
- System principle (material handling), 501–502
- System safety, 700–708
 - analysis methods, 700–702
 - fault tree technique, 702

- System safety (*continued*)
 preparation/review of procedures criteria,
 702–707
 risk assessment process, 707–708
 System safety analysis, 700
 System–subsystem integration, 701
- T**
- Tactical level (manufacturing systems), 186, 188
 Taguchi, Genichi, 649
 Takt time, 624
 TAM, *see* Thermally assisted machining
 Tardy jobs, minimizing number of, 89–90
 Taxes, 600, 801–802
 Taylor, Frederick, 6
 TBC (time-based competition), 107
 TCM (thermochemical machining), 179–180
 Teams, 8–10, 16, 530, 565, 570–574
 Teambuilding, 571, 576
 Technical contradictions (TRIZ), 366
 Technical expertise, building, 575
 Technique for human error prediction (THERP),
 700
 Technological development. *See also* Process
 technology
 history of, 397–398
 workplace and economic changes from, 561
 Technology-based projects, evaluating/selecting,
see Project evaluation and selection
 Telecommunication industry standard (TL 9000),
 674
 TEM (thermal energy method), 179
 Template machining (gears), 145
 Temporary corrosion protection, 232
 Tensile loads, theoretical behavior of bolted/riveted
 joints under, 272–276
 TFM, *see* Total form machining
 Theory to the solution of inventive problems, 362
 Thermal deburring, 179
 Thermal effects (O-rings), 291
 Thermal energy method (TEM), 179
 Thermal inspection, 468–469
 Thermally assisted machining (TAM), 159, 160,
 164
 Thermal paints, 469
 Thermal testing, 444, 469
 Thermochemical machining (TCM), 179–180
 Thermoforming (plastics), 226
 Thermoplastics, 230
 Thermoplastic resins, 47
 Thermosetting resins, 47
 THERP (technique for human error prediction),
 700
 Thornton structure zone diagram, 239
 Thread cutting and forming, 146–148
 Thread rolling, 148, 210
 Thread run-out, 281
 Thread stress distribution, 281
 Three-part assemblies (mechanical fasteners),
 260–261
 Tier II suppliers, 30
 Tier I suppliers, 30
 Time-based competition (TBC), 107
 Time bucket, 78
 Time-series analysis, 55, 60–62
 Time-series forecast, 55
 Time value of money, 583–584
 TL 9000: telecom quality management system,
 674
 Tolerance(s):
 of assembled products, 331
 relaxation of, 405
 statistical quality control, 325, 331
 Tooling modification, 405
 Tool life, 121–124
 Tort claims, 751, 753
 Total form machining (TFM), 159, 161, 164
 Total productive maintenance (TPM), 623
 Total quality management (TQM), 635–664
 and Deming Prize, 678
 DMADV, 636–637, 656–664
 analyze phase, 660–661
 define phase, 656–657
 design phase, 661–663
 measure phase, 657–660
 verification and validation phase, 663–664
 DMAIC, 636–656
 analyze phase, 643–649
 control phase, 653–656
 define phase, 637–639
 improve/innovate phase, 649–653
 measure phase, 639–643
 Toxic Substances Control Act (TSCA, 1976),
 695
 Toyota, Eiji, 618
 Toyota, Kiichiro, 618
 Toyota, Sakichi, 617
 Toyota Automatic Loom Works, 617
 Toyota, 21
 Toyota Motor Company, 617–618, 623
 Toyota Production System (TPS), 15, 21, 618.
See also Lean management
 TPM (total productive maintenance), 623
 TPS, *see* Toyota Production System
 TQM, *see* Total quality management
 TRA, *see* Tuszynski's relational algorithm
 Trade secrets, 754

- TRAIL, 817
- Training:
- for improvement systems implementation, 22–23
 - for PPE use, 725
 - safety, 735–745
 - automated external defibrillator, 735
 - driver, 735–736
 - environmental risk, 736–739
 - fire protection, 739
 - first-aid, 739
 - hazard recognition, 739–740
 - HAZWOPER, 740
 - job hazard analysis, 742–744
 - lockout box, 740–741
 - machine guard, 741
 - management oversight of, 744–745
 - personal protection equipment/physical protection, 742
 - sources/types of training materials, 745
 - specialized courses, 735–742
 - Transit time instruments (bolts), 279–280
 - Transmission (sound), 458–459
 - Transportation hazard analysis, 701
 - Transporters, environmentally benign manufacturing and, 31, 32
 - Traveling wire EDM, 175
 - Tree diagram, 629, 631
 - Trend, 55
 - Trend movements (time series), 60
 - Trepanning, 139
 - Trimming (metals), 213, 215
 - Triode sputtering (PVD), 244–245
 - TRIZ, 361–388
 - analytical tools, 375–376
 - ARIZ algorithm for inventive problem solving, 383–388
 - caveat for, 388
 - steps in, 384–385
 - Class 4 measurement and detection standards, 379–383
 - contradiction matrix, 371–373
 - contradictions principle in, 365–367
 - ideality principle in, 364–365
 - laws of systems evolution, 374–375
 - maximal use of resources principle in, 367
 - origins of, 362–364
 - physical contradictions, 371, 373
 - problems without contradictions, 376–378
 - requirements for inventive solutions in, 370
 - scientific approach of, 367–370
 - solution by abstraction using, 368–370
 - Su-Field models, 376, 378–379
 - tools of, 370–376
 - analytical, 375–376
 - contradiction matrix, 371–373
 - laws of systems evolution, 374–375
 - physical contradictions, 371, 373
 - Su-Field models, 376
 - Trucks, 504
 - TSCA (Toxic Substances Control Act, 1976), 695
 - Tube spinning, 206
 - Tumbling, 229
 - Turbine engine seals, 294, 296, 317, 319
 - Turning:
 - defined, 129
 - electrochemical, 170, 171
 - primary factors, 129, 130
 - tool wear factors, 122
 - Turning machines, 129–133
 - “Turtle diagram” (ISO 9001 certification), 671–672
 - Tuszynski’s relational algorithm (TRA), 406–432
 - across engineering functionalities, 408
 - correlation charts, 412, 415, 416
 - definitions, 435–440
 - fixes, 424–428
 - foundation of, 409
 - generating input data for, 411–413
 - graphical illustration of, 409, 410
 - history of, 406–407
 - material selection, 428–432
 - operating point adjustments, 411
 - process conclusions for, 409, 410
 - region of conformance, 414, 416
 - relationship conditions in, 414, 416–425
 - selecting predictors, 412, 414, 415
 - as single-degree-of-freedom system, 410–411
 - usefulness of, 409
 - what it is not, 407
- U**
- UAM (ultrasonic abrasive machining), 159
 - U bending, 212
 - UBM (unbalanced magnetron) effect, 245–247
 - u* charts, 334–335
 - Ultimate tensile strength (bolts), 270
 - Ultrasonic abrasive machining (UAM), 159
 - Ultrasonic inspection, 456–465
 - bond testers, 464–465
 - capabilities, 444
 - inspection process, 461–464
 - reflection and transmission of sound, 458–459

Ultrasonic inspection (*continued*)
 refraction of sound, 459–461
 sound waves, 457–458
 ultrasonic properties of common materials,
 476–493

Ultrasonic machining (USM), 157, 159, 161, 163,
 165

Ultrasonic properties of common materials,
 476–493

Unbalanced magnetron (UBM) effect, 245–247

Unforeseeable misuse (product liability), 762

Uniform series (cash flow), 588

Unions, 7

U.S. Department of Defense (DoD), 336, 668, 675,
 676

U.S. Environmental Protection Agency (EPA),
 694–697

Clean Air Act, 696

Clean Water Act, 694

Comprehensive Environmental Response,
 Compensation, and Liability Act,
 695–696

Federal Insecticide, Fungicide, and Rodenticide
 Act, 697

National Environmental Policy Act, 694

Oil Pollution Act, 696

Pollution Protection Act, 697

Resource Conservation and Recovery Act,
 694–695

Superfund Amendments and Reauthorization
 Act, 696

Toxic Substances Control Act, 695

U.S. Missile Defense Agency (MDA), 674–675

U.S. National Aeronautics and Space
 Administration (NASA), 674–675,
 677

U.S. Patent and Trademark Office (USPTO),
 779, 788, 790, 791, 813, 817.
See also Patents

U.S. state quality awards, 684, 687

Unit load principle (material handling), 500

Unknown demand (inventory), 64

Unsafe conditions. *See also* Safety engineering
 checklist for mechanical or physical facilities,
 734–735
 eliminating, 729–734

Upset forging (metals), 201

URL, 809

USM, *see* Ultrasonic machining

USPTO, *see* U.S. Patent and Trademark Office

Utility (patentability requirement), 782

Utility application (patents), 775, 776

Utility patents, 774, 775, 778

V

Vacuum metallizing, 231–232

Validating improvements, 650

Validation tests, 664

Value engineering, 16

Value stream, 618–619

Vapor baths (cleaning), 228–229

Variable characteristics, 400

Varnishes, 231

V bending, 211, 212

Ventilation, as safety control, 719

Verification and validation phase (DMADV),
 663–664

Verification and validation plans (DMADV), 663

Verification tests, 663–664

Vertical integration (manufacturing systems), 20

Vertical organizations, 7

Vibration hazard, 713

Viruses (computer), 809, 811–812

Virus protection, 811

Visual factory, 622

Vitreous enamels, 231

Voice of the customer (VOC), 657

Volatile organic compound (VOC) emissions, 302

W

Warehouse material handling devices, 505

Warehousing, 501, 502, 505, 509–510

Warnings:
 continuing duty to warn, 768–770
 of product defects, 761–762, 764

Waste:
 in environmentally conscious manufacturing,
 184
 Hazardous and Solid Waste Amendments, 695
 HAZWOPER training, 740
 from injection molding, 50
 in lean management, 618–620
 Resource Conservation and Recovery Act,
 694–695
 from sand casting, 36

Water-jet machining (WJM), 159, 163–165

Waterman, Robert H., 624

Watermark Designs, 22

Weighted mean flow time (single-machine
 problem), 89

Weighted moving average, 56–57

WELDASEARCH, 815

Welding:
 arc-welding, 41–43
 electron beam, 44
 environmentally benign manufacturing, 41–44

gas-flame, 41, 42
pipe, 206–207
resistance-welding, 41, 43
“Whack-A-Mole” approach, 646
Wind chill index, 738–739
Wire brushing, 229
Wire cutting, electrical discharge, 174–175
Withdrawal kanban, 105
WJM (water-jet machining), 159, 163–165
Work element, 98
Work element time, 98
“Worker Attitudes and Perceptions of Safety”
(ReVelle and Boulton), 692, 693
Workforce:
attitudes and perception of safety in, 692, 693
reductions of, 16–17
social engineering, 15–17
Workforce costs, 74
Work management, 562

Work principle (material handling), 499
Work process (term), 574
Workstation, 98
WPINDEX, 815

X

\bar{X} chart, 326–331
X-ray imaging technologies, 448. *See also*
Radiography testing (RT)

Y

Yield point (bolts), 270

Z

Zlotin, Boris, 371
Zotero, 819

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.