HANDBOOK OF MILITARY INDUSTRIAL ENGINEERING

EDITED BY

ADEDEJI B. BADIRU

MARLIN U. THOMAS

CRC Press
Taylor & Francis Group

HANDBOOK OF
MILITARY INDUSTRIAL ENGINEERING

Industrial Innovation Series

Series Editor

Adedeji B. Badiru

Department of Systems and Engineering Management
Air Force Institute of Technology (AFIT) – Dayton, Ohio

PUBLISHED TITLES

Computational Economic Analysis for Engineering and Industry
Adedeji B. Badiru & Olufemi A. Omitaomu

Handbook of Industrial and Systems Engineering
Adedeji B. Badiru

Handbook of Military Industrial Engineering
Adedeji B.Badiru & Marlin U. Thomas

Industrial Project Management: Concepts, Tools, and Techniques
Adedeji B. Badiru, Abidemi Badiru, and Adetokunboh Badiru

Knowledge Discovery from Sensor Data
Auroop R. Ganguly, João Gama, Olufemi A. Omitaomu, Mohamed Medhat Gaber, and Ranga Raju Vatsavai

Systems Thinking: Coping with 21st Century Problems
John Turner Boardman & Brian J. Sauser

Techonomics: The Theory of Industrial Evolution
H. Lee Martin

Triple C Model of Project Management: Communication, Cooperation, Coordination
Adedeji B. Badiru

FORTHCOMING TITLES

Beyond Lean: Elements of a Successful Implementation
Rupy (Rapinder) Sawhney

Handbook of Industrial Engineerng Calculations and Practice
Adedeji B. Badiru & Olufemi A. Omitaomu

Industrial Control Systems: Mathematical and Statistical Models and Techniques
Adedeji B. Badiru, Oye Ibidapo-Obe, & Babatunde J. Ayeni

Modern Construction: Productive and Lean Practices
Lincoln Harding Forbes

Project Management: Systems, Principles, and Applications
Adedeji B. Badiru

Research Project Management
Adedeji B. Badiru

Statistical Techniques for Project Control
Adedeji B. Badiru

STEP Project Management: Guide for Science, Technology, and Engineering Projects
Adedeji B. Badiru

Technology Transfer and Commercialization of Environmental Remediation Technology
Mark N. Goltz

HANDBOOK OF
MILITARY INDUSTRIAL ENGINEERING

EDITED BY
ADEDEJI B. BADIRU
MARLIN U. THOMAS

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-6628-9 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Handbook of military industrial engineering / editors, Adedeji B. Badiru, Marlin U. Thomas.
p. cm. -- (Industrial innovation series)
"A CRC title."
Includes bibliographical references and index.
ISBN 978-1-4200-6628-9 (hardcover : alk. paper)
1. Military engineering--Handbooks, manuals, etc. 2. Industrial engineering--Handbooks, manuals, etc. I. Badiru, Adedeji Bodunde, 1952- II. Thomas, Marlin U., 1942- III. Title. IV. Series.

UG150.H46 2009
355.4--dc22 2009001090

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To our families,
who stood by us through it all.

Contents

Preface

Recent global events have necessitated the need to look at military operations more critically in terms of process design, planning, management, improvement, and control. Industrial engineering offers one proven approach to achieving that goal. Industrial engineering has been applied to military operations for many decades. Unfortunately, until now, there has not been a focused collection of the applications for ease of reference. The *Handbook of Military Industrial Engineering* presents a compilation of the fundamental tools of industrial engineering techniques with a focus on military applications. Examples of the roles that industrial engineering plays in military operations can be seen in many present operational strategies of the military. Industrial engineering is well versed and appropriately positioned to create, adapt, utilize, and disseminate new knowledge and tools for direct application to military operations. The versatility of industrial engineering has been demonstrated again and again over the years. It is through the application of industrial engineering principles, tools, and techniques that many operational improvements have been achieved in many organizations.

The chapters in this handbook are contributed by well-known industrial engineering authors, researchers, educators, and practitioners. The contents of the book will help military organizations to effectively evaluate operational constraints of time, cost, and performance. The techniques and applications contained in the book cut across all military services: Air Force, Navy, Army, Marines corps.

The utility of the handbook is not limited to one national focus. Mutual applications can be found in friendly and cooperating nations. In military organizations, workers need to know which tools and techniques to use for the various operational challenges that they face in day-to-day functional responsibilities. In operational crises, industrial engineering techniques can be used to identify sources of problems and to make operational corrections as well as respond to emergencies efficiently and effectively. Industrial engineering research and development can lead to the identification of the most effective ways to use human resources, tools, and work processes. Thus, this handbook has potential utility in nonmilitary applications.

Part I of the handbook presents an Executive Summary of military applications of industrial engineering. Part II consists of chapters on Modeling and Optimization. Part III consists of chapters on Reliability and Maintenance. Part IV consists of chapters on Contingency Planning and Logistics. Part V consists of chapters on Supply Chain and Decision Making. Part VI consists of chapters on Human Factors and Ergonomics. Part VII consists of chapters on Management and Process Improvement. The handbook is expected to be a valuable reference material for students, educators, researchers, policy makers, and practitioners.

Adedeji Badiru
Marlin U. Thomas
Air Force Institute of Technology

Acknowledgments

We thank all our colleagues, friends, and family who supported us and found reasons to bear with us throughout the long journey of putting together this monumental compilation. We appreciate the intellectually rich environment of the Air Force Institute of Technology, which provided the motivation for initiating this handbook project. We also thank Iswat Badiru for her typing and administrative support for the project. She diligently retyped and reformatted many of the initial manuscript submissions including several graphics contained in the handbook. Without her dedicated support, this project could not have been completed within the desired schedule, cost, and quality metrics.

Editors

Adedeji Badiru is a professor and department head at the Department of Systems and Engineering Management at the Air Force Institute of Technology, Dayton, OH. He was previously professor and department head of Industrial and Information Engineering at the University of Tennessee in Knoxville. Prior to that, he was professor of industrial engineering and Dean of University College at the University of Oklahoma. He is a registered professional engineer, a fellow of the Institute of Industrial Engineers, and a fellow of the Nigerian Academy of Engineering. He holds BS in Industrial Engineering, MS in Mathematics, and MS in Industrial Engineering from Tennessee Technological University, and PhD in Industrial Engineering from the University of Central Florida. His areas of interest include mathematical modeling, project modeling and analysis, economic analysis, and productivity analysis and improvement. He is the author of several books and technical journal articles. He is a member of several professional associations including Institute of Industrial Engineers (IIE), Institute of Electrical and Electronics Engineers (IEEE), American Society of Engineering Management (ASEM) Society of Manufacturing Engineers (SME), Institute for Operations Research and Management Science (INFORMS), American Society for Engineering Education (ASEE), New York Academy of Science (NYAS), and Project Management Institute (PMI).

He has served as a consultant to several organizations around the world including Russia, Mexico, Taiwan, Nigeria, and Ghana. He has conducted customized training workshops for numerous organizations including Sony, AT&T, Seagate Technology, US Air Force, Oklahoma Gas and Electric, Oklahoma Asphalt Pavement Association, Hitachi, Nigeria National Petroleum Corporation, and ExxonMobil. He has won several awards for his teaching, research, publications, administration, and professional accomplishments. He holds a leadership certificate from the University Tennessee Leadership Institute. Professor Badiru has served as a technical project reviewer, curriculum reviewer, and proposal reviewer for several organizations including The Third-World Network of Scientific Organizations, Italy, National Science Foundation, National Research Council, and the American Council on Education. He is on the editorial and review boards of several technical journals and book publishers. Professor Badiru has also served as an Industrial Development Consultant to the United Nations Development Program.

Marlin U. Thomas is dean of the Graduate School of Engineering and Management at the Air Force Institute of Technology. He was previously a professor and past head of the School of Industrial Engineering at Purdue University where he has also served as director of the Institute for Interdisciplinary Engineering Studies. He received his BSE (1967) from the University of Michigan-Dearborn, and MSE (1968) and PhD (1971) from the University of Michigan, all in industrial engineering. He is a registered professional engineer in Michigan, with engineering R&D and product development experience with Owens-Illinois, Inc., and as a manager of vehicle reliability planning and analysis with Chrysler Corporation. He has also served as a program director for the National Science Foundation for operations research and production systems. His teaching and research areas include stochastic modeling, reliability and logistics systems, with an emphasis on contingency operations. He has authored or coauthored over 75 articles and technical papers that have appeared in *IIE Transactions, Operations Research, Management Science, IEEE Transactions on Reliability*, and others and delivered over 100 presentations and lectures at major conferences and universities.

Professor Thomas has held a number of leadership positions including past-president of the Institute of Industrial Engineers; secretary of the Operations Research Society of America; department editor, *IIE Transactions*; area editor, *Operations Research*; and member of the Army Science Board. He has received several awards for his professional accomplishments including the Institute of Industrial Engineering's highest honor, the Frank and Lillian Gilbreth Award. He is a member of the Military Operations Research Society, American Society for Engineering Education, American Indian Science and Engineering Society; and fellow of the American Society for Quality, the Institute of Industrial Engineers, and the Institute for Operations Research and Management Sciences. He is also a retired captain in Civil Engineer Corps for the US Navy Reserve.

Contributors

Karim Abdel-Malek
Departments of Biomedical and Mechanical Engineering
The University of Iowa
Iowa City, Iowa

Ravindra K. Ahuja
Department of Industrial and Systems Engineering
University of Florida
Gainesville, Florida

Charles H. Aikens
Department of Industrial and Information Engineering
University of Tennessee
Knoxville, Tennessee

Anthony F. Antoline
Maintenance Studies Branch
Air Force Logistics Management Agency
Montgomery, Alabama

Achilles A. Armenakis
Department of Business Management
Auburn University
Auburn, Alabama

Jasbir Arora
Department of Civil and Environmental Engineering/ Department of Mechanical and Industrial Engineering
The University of Iowa
Iowa City, Iowa

Michael J. Artelli
Directorate of Analyses and Lessons Learned
Air Combat Command
Langley Air Force Base
Newport News, Virginia

Adedeji B. Badiru
Department of Systems and Engineering Management
Air Force Institute of Technology
Wright Patterson Air Force Base
Dayton, Ohio

J. Wesley Barnes
Graduate Program in Operations Research and Industrial Engineering
The University of Texas at Austin
Austin, Texas

Rajan Batta
Department of Industrial and Systems Engineering
University at Buffalo
Buffalo, New York

Dan O. Bausch
Operations Research Department
Naval Postgraduate School
Monterey, California

Misty Blue
Department of Biomedical, Industrial and Human Factors Engineering
Wright State University
Dayton, Ohio

Jesse W. Brogan
Walter Reed Army Institute of Research
Silver Spring, Maryland

Gerald G. Brown
Operations Research Department
Naval Postgraduate School
Monterey, California

W. Peter Cherry
SAIC: Future Combat Systems
Vienna, Virginia

Jeffery K. Cochran
Department of Operational Sciences
Air Force Institute of Technology
Wright Patterson Air Force Base
Dayton, Ohio

Darwin J. Davis
Department of Business Administration
University of Delaware
Newark, Delaware

Richard F. Deckro
Department of Operational Sciences
Air Force Institute of Technology
Wright Patterson Air Force Base
Dayton, Ohio

Frank Dement
Robins Air Force Base
Georgia

Hubert S. Feild
Shool of Business
Auburn University
Auburn, Alabama

Bobbie L. Foote
United States Military Academy
West Point, New York

Michael L. Fredley
National Reconnaissance Office
Chantilly, Virginia

George R. Freeman
Air Force Center for Systems Engineering
Air Force Institute of Technology
Wright Patterson Air Force Base
Dayton, Ohio

Simon R. Goerger
National War College
Fort McNair
Washington, DC

Roger D. Golden
Maintenance Studies Branch
Air Force Logistics Management Agency
Montgomery, Alabama

Stanley G. Harris
Department of Business Management
Auburn University
Auburn, Alabama

Raymond R. Hill
Department of Operational Sciences
Air Force Institute of Technology
Wright Patterson Air Force Base
Dayton, Ohio

Thom J. Hodgson
Fitts Department of Industrial and Systems Engineering
North Carolina State University
Raleigh, North Carolina

Daniel T. Holt
Department of Systems and Engineering Management
Air Force Institute of Technology
Dayton, Ohio

Jeremy A. Howe
Maintenance Studies Branch
Air Force Logistics Management Agency
Montgomery, Alabama

Reiner K. Huber
University of the Federal Armed Forces
Munich, Germany

Danny R. Hundley
Operations Research Department
Naval Postgraduate School
Monterey, California

Anota O. Ijaduola
Department of Physics
Monmouth College
Monmouth, Illinois

Mohamad Y. Jaber
Department of Mechanical and Industrial Engineering
Ryerson University
Toronto, Ontario, Canada

Jack A. Jackson
Institute for Defense Analyses
Alexandria, Virginia

Krishna C. Jha
Innovative Scheduling, Inc.
Gainesville, Florida

Doug Keene
Robins Air Force Base
Georgia

Hemant V. Kher
Department of Business Administration
University of Delaware
Newark, Delaware

Kathrin Klamroth
Department of Mathematics
University of Wuppertal
Wuppertal, Germany

Arvind Kumar
Innovative Scheduling, Inc.
Gainesville, Florida

Way Kuo
College of Engineering
City University of Hong Kong
Hong Kong, People's Republic of China

Douglas M. Lambert
Fisher College of Business
The Ohio State University
Columbus, Ohio

Mark A. Lawley
Department of Industrial Engineering
Purdue University
West Lafayette, Indiana

Youngho Lee
Division of Information Management Engineering
Korea University
Seoul, Republic of Korea

Erjen Lefeber
Department of Mechanical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands

Lawrence M. Leemis
Department of Mathematics
The College of William and Mary
Williamsburg, Virginia

Theodore P. Lewis
Air Force Material Command Headquarters
Wright Patterson Air Force Base
Dayton, Ohio

Tom Marechal
Department of Mechanical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands

Timothy Marler
Center for Computer Aided Design
The University of Iowa
Iowa City, Iowa

Barbra Melendez
Department of Mathematics
United States Military Academy
West Point, New York

James T. Moore
Department of Operational Sciences
Air Force Institute of Technology
Wright Patterson Air Force Base
Dayton, Ohio

Rakesh Nagi
Department of Industrial and Systems Engineering
University at Buffalo
Buffalo, New York

Gary O'Neill
Robins Air Force Base
Georgia

James B. Orlin
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, Massachusetts

Scotty A. Pendley
Maintenance Studies Branch
Air Force Logistics Management Agency
Montgomery, Alabama

Barbara Pfeiffer
Department of Mathematics
University of Wuppertal
Wuppertal, Germany

Edward A. Pohl
Department of Industrial Engineering
University of Arkansas
Fayetteville, Arkansas

Stephen H. Rapp
Operations Research Department
Naval Postgraduate School
Monterey, California

A. Ravi Ravindran
Department of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, Pennsylvania

Richard E. Rosenthal
Operations Research Department
Naval Postgraduate School
Monterey, California

Hanif D. Sherali
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Alice E. Smith
Department of Industrial and Systems Engineering
Auburn University
Auburn, Alabama

Jeffrey S. Smith
Department of Industrial and Systems Engineering
Auburn University
Auburn, Alabama

W. Owen Spivey
Combat Development Engineering Field Office
Fort Eustis, Virginia

Benjamin A. Thoele
Maintenance Studies Branch
Air Force Logistics Management Agency
Montgomery, Alabama

Kristin A. Thoney
College of Textiles
North Carolina State University
Raleigh, North Carolina

Marlin U. Thomas
Graduate School of Engineering and Management
Air Force Institute of Technology
Wright Patterson Air Force Base
Dayton, Ohio

Tim E. Trainor
Department of Systems Engineering
United States Military Academy
West Point, New York

Volkan Ustun
Department of Industrial and Systems Engineering
Auburn University
Auburn, Alabama

John C. Van Hove
Studies and Analysis Squadron
Air Education and Training Command
Randolph Air Force Base
San Antonio, Texas

Vijay Wadhwa
Department of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, Pennsylvania

Rui Wan
Department of Industrial and Information Engineering
University of Tennessee
Knoxville, Tennessee

Michael K. Williams
Transportation Engineering Agency
JDPAC & SDDCTEA
Scott Air Force Base
Illinois

Jingzhou (James) Yang
Department of Mechanical Engineering
Texas Tech University
Lubbock, Texas

I

Executive Summary

1

Executive Summary: Handbook of Military Industrial Engineering

Adedeji B. Badiru
Air Force Institute of Technology

Industrial engineering (IE)—the design and development of integrated systems of people, machines, and information resources for producing products and services.

- Operations research and industrial engineering methods originated out of needs for conserving critical resources during the war.
- Following World War II there was a follow-up need for optimization and improvement methods for developing businesses and industry.
- Major military applications have dealt in areas of effective methods and techniques for managing and planning resources for providing logistics and support for training and operations.
- IE philosophy—"there is always a better way".

1.1 Industrial Engineering

1.1.1 Definition of an Industrial Engineer

> "Industrial engineer—*one who is concerned with the design, installation, and improvement of integrated systems of people, materials, information, equipment, and energy by drawing upon specialized knowledge and skills in the mathematical, physical, and social sciences, together with the principles and methods of engineering analysis and design to specify, predict, and evaluate the results to be obtained from such systems.*" (From Institute of Industrial Engineers, www.iienet.org)

1.1.2 Philosophy

The comprehensive definition of industrial engineering embodies the various aspects of the diversity and versatility of the profession. For several decades, the military had called upon the discipline of industrial engineering to achieve program effectiveness and operational efficiencies. Industrial engineering is versatile, flexible, adaptive, and diverse. It can be seen from the definition that a systems orientation permeates the work of industrial engineers. This is particularly of interest to the military because military operations and functions are constituted by linking systems.

1.2 IE/OR and National Defense

- Evolution of optimization and methods for achieving efficient utilization of resources.
- Logistics is a major issue for military planning and operations.

1.3 Scope of the Book

The *Handbook of Military Industrial Engineering* covers principles, tools, and modeling techniques of industrial engineering with specific and direct applications to military systems. The diversity of industrial engineering cuts across several disciplines and applications. The scope of the handbook covers the following major topical areas:

1. Modeling and optimization.
2. Reliability and maintenance.
3. Contingency planning and logistics.
4. Supply chain and decision making.
5. Human factors and ergonomics.
6. Management and process improvement.

1.4 Summary of Contents

Chapter 1 focuses on using a highly detailed simulation model to create a physical security system to detect intrusions by human agents in a building, thereby minimizing the expected loss of such intrusions. Security consists of human guards and fixed security cameras. The problem is represented as a binary optimization problem and a new heuristic is proposed to perform the security configuration optimization, the Hybrid Genetic Ant Optimization (HGAO). This heuristic combines a genetic algorithm and an ant colony optimization. HGAO is compared with a genetic algorithm for two security optimization scenarios.

Chapter 2 discusses the problem of scheduling a set of target illuminators (passive homing devices) to strike a set of targets using surface-to-air missiles in a naval battle-group antiair warfare scenario. This problem is conceptualized as a parallel (unrelated) machine scheduling problem of minimizing the total weighted flow time, subject to time-window job availability and machine downtime side-constraints. The chapter presents a zero-one integer programming formulation for this problem and devises some classes of strong valid inequalities to enhance the model representation. By exploiting the special structures inherent in the formulation, effective heuristics are developed that provide near-optimal solutions in real time; this was an essential element for use by the Navy. A major computational bottleneck is the solution of the underlying linear programming relaxation because of the extremely high degree of degeneracy inherent in the formulation. In order to overcome this difficulty, a Lagrangian dual formulation is employed to generate lower and upper bounds, which are then embedded in a branch-and-bound algorithm. Computational results are presented using suitable realistic data.

Chapter 3 presents a study of the problem of scheduling a set of target illuminators (passive homing devices) to strike a set of targets using surface-to-air missiles in a naval battle-group antiair warfare scenario. This problem is conceptualized as a parallel (unrelated) machine scheduling problem of minimizing the total weighted flow time, subject to time-window job availability and machine downtime side-constraints. A zero-one integer programming formulation is presented for this problem and some classes of strong valid inequalities to enhance the model representation are devised. By exploiting the special structures inherent in the formulation, effective heuristics are developed that provide near-optimal solutions in real time; this was an essential element for use by the Navy. A major computational bottleneck is the solution of the underlying linear programming relaxation because of the extremely high degree of degeneracy inherent in the formulation. In order to overcome this difficulty, a Lagrangian dual formulation is employed to generate lower and upper bounds, which are then embedded in a branch-and-bound algorithm. Computational results are presented using suitable realistic data.

Chapter 4 presents multiple criteria optimization models for supplier selection. Supplier selection process is difficult because the criteria for selecting suppliers could be conflicting. Supplier selection is a multiple criteria optimization problem that requires trade-off among different qualitative and quantitative factors to find the best set of suppliers. For example, the supplier with the lowest unit price may also have the lowest quality. The problem is also complicated by the fact that several conflicting criteria must be considered in the decision-making process. Most of the times buyers have to choose among a set of suppliers by using some predetermined criteria such as, quality, reliability, technical capability, lead-times etc., even before building long-term relationships. To accomplish these goals, two basic and interrelated decisions must be made by a firm. The firm must decide which suppliers to do business with and how much to order from each supplier. A supply chain consists of a connected set of activities concerned with planning, coordinating and controlling materials, parts and finished good from supplier to customer. The contribution of the purchasing function to the profitability of the supply chain has assumed greater proportions in recent years; one of the most critical functions of purchasing is selection of suppliers. For most manufacturing firms, the purchasing of raw material and component parts from suppliers constitutes a major expense.

Chapter 5 focuses on path planning for unmanned aerial vehicles (UAVs) in the presence of threat zones is traditionally treated in the literature by avoidance of such areas. However, when the UAV must reach the target within a specified time, threat zones might need to be entered. To address such a situation, the authors develop a probabilistic modeling framework in this chapter for UAV routing in a continuous setting. The chapter considers several problems using this modeling construct: (a) computation of a minimum risk origin–destination path in the presence of threat zones (nested convex polygonal and circular); (b) exploration of the trade-off between flying time versus risk using bi-criteria optimization; (c) expected time until the first detection as a tiebreaker; and (d) consideration of the effect of a delay between detection and fatal attack. The chapter also presents examples to illustrate the results.

Chapter 6 presents a technique for modeling the end-to-end military transportation problem. Moving people, equipment, and supplies to support military endeavors has been accomplished since the dawn of recorded history. Whenever peoples in distant lands have been engaged in conflicts with each other, the ability to move people (warriors) and their equipment and supplies successfully has been a key to the outcome of the conflict. The opponent who has to move the furthest to engage in conflict has found the accomplishment of the deployment and distribution task critical. In order to succeed, the "right things must be at the right place at the right time". Historical records are filled with examples of where the winner has done this successfully and the loser has not. A few notable examples are offered to support this observation. Alexander the Great conquered a large portion of the known world by moving large numbers of soldiers across large distances, living off the land, and incorporating conquered peoples into his army (thus extending his support and supply lines). Hannibal moved armies and their tools of war over mountains to attack Rome. Napoleon's armies were able to cover large distances and secure large holdings because of their ability to carry supplies and supplement their supplies from the conquered economies. Skipping ahead to more recent times, German armies moved at great speed during World War II because of their superior support mechanism. Yet, when these vast armies invaded Russia, they could not win. One of the major reasons for failure was their lack of adequate winter clothing. Thus, modeling the end-to-end military transportation problem from a holistic approach is very essential for military success.

Chapter 7 presents the weapon—target assignment (WTA) problem, which is a fundamental problem arising in defense-related applications of operations research. This problem consists of optimally assigning n weapons to m targets so that the total expected survival value of the targets after all the engagements is minimal. The WTA problem can be formulated as a nonlinear integer programming problem and is known to be NP-complete. No exact methods exist for the WTA problem that can solve even small size problems (for example, with 20 weapons and 20 targets). Though several heuristic methods have been proposed to solve the WTA problem, due to the absence of exact methods, no estimates are available on the quality of solutions produced by such heuristics. This chapter suggests integer programming and network-flow-based lower bounding methods which is obtained using a branch and bound algorithm for the WTA problem. The chapter also proposes a network-flow-based construction heuristic and a very large-scale neighborhood (VLSN) search algorithm. The chapter includes computational results of the algorithms which indicate that they can solve moderately large instances (up to 80 weapons and 80 targets) of the WTA problem optimally and obtain almost optimal solutions of fairly large instances (up to 200 weapons and 200 targets) within a few seconds.

Chapter 8 presents an optimization model for military budget allocation and capital rationing. Protection of national interests is accomplished over several capabilities. These capabilities must all be funded at appropriate levels and on a relative basis in order to achieve desired outcomes. In an era of diminishing budgetary resources, strategic allocation of limited resources must be exercised. It has been shown that heuristic approaches to budget allocation do not work effectively for complex decision problems. Military operations constitute such a complex decision problem. As warfare becomes more network-centric, strategies must be developed to ensure that each element of the composite network receives mission-appropriate resource allocation. This is accomplished through optimization of budget allocation and capital rationing. This chapter presents an example of a linear programming model developed for budget allocation by the Commander of Navy Installations (CNI). The modeling approach can be adapted to other military or commercial applications.

Chapter 9 focuses on an overview of meta-heuristics and their use in military modeling. The military establishment uses many types of models during planning and execution of military operations. Optimization models have always been an important class of models. However, as military systems, the scenarios in which those military systems are employed, and the impact of operations involving those military systems become more complex, classic optimization techniques based on mathematical programming can become impractical. Advances in computing capabilities have helped advance the use of heuristics to quickly find reasonable solutions to complex optimization problems. This chapter provides an overview of the more popular, and powerful, modern heuristics, most of which have analogies to natural systems. A representative survey is provided of how these heuristics have been applied to specific classes of military optimization applications.

Chapter 10 discusses recent advances in optimal reliability allocation. Reliability has become a greater concern in recent years because high-tech industrial processes with ever increasing levels of sophistication comprise most engineering systems today. To keep pace with this rapidly developing field, this chapter provides a broad overview of recent research on reliability optimization problems and their solution methodologies. In particular, the following issues are addressed:

(a) UGF-based optimal multi-state system design.
(b) Percentile life employed as a system performance measure.
(c) Multi-objective optimization of reliability systems, especially with uncertain component reliability estimations.
(d) Innovation and improvement in traditional reliability optimization problems, such as fault-tolerance mechanism and cold-standby redundancy involved system design.

New developments in optimization techniques are also emphasized in this chapter, especially the methods of ant colony optimization and hybrid optimization. The interesting problems that are reviewed here are deserving of more attention in the literature. To that end, the chapter concludes with a discussion of future challenges related to reliability optimization.

Chapter 11 addresses the problem of determining lower confidence bounds on the system reliability for a coherent system of k components using failure data. The methodology is developed under the assumption that the components fail independently. That is, there are no common-cause failures. Binary failure data are collected for each of the independent components in a coherent system. Bootstrapping is used to determine lower confidence bound on the system reliability. When a component with perfect test results is encountered, a beta prior distribution is used to avoid an overly optimistic lower bound.

Chapter 12 presents a methodology for assessing the reliability of a contingency logistics network. A method is presented for determining the reliability of a logistics network for supporting contingency operations. The method is based on the interference of random load and capacity of the distribution links of a supply chain. Conditions of risk are considered using various assumed distributions, and uncertainty based on maximum entropy distributions. Examples are provided to illustrate these conditions. Contingencies are unexpected crises events that create a major threat to the safety and security of a population. They range from military conflicts that require engagement with hostile forces, police actions for civil disorder, to humanitarian relief of victims from disasters such as earthquakes, hurricanes, floods, and related catastrophes. Hurricane Mitchell, the overthrow of the government of Somalia, and the 1990 invasion of Kuwait are examples of crises events that resulted in major contingency operations. Contingency events trigger immediate need for logistics support functions to organize and mobilize people, equipment, materials and supplies for carrying out the emergency operations. These operations are generally initiated from existing plans by the Federal Emergency Management Agency (FEMA) for disasters that occur within the US continent, or the Department of Defense for national contingencies outside the continent.

Chapter 13 presents a technique for computing small-fleet aircraft availabilities including redundancy and spares. Logistics support is a key element of aircraft transportation systems. The chapter is concerned with the impact of aircraft spares provisioning decisions on the availability of aircraft. Spares provisioning in this context is complicated by the fact that spares may be shared across aircraft and that aircraft may have redundant systems. In addition, decisions concerning aircraft spares support require a rapid response for safety reasons. Analytical models have proven to provide a quicker response time than corresponding simulation models. There is an existing analytical model that includes the effect of redundancy and spares, but the underlying assumption is that a large number of aircraft are being modeled. In many applications, predictions of the number of times an aircraft can fly each day and the number of aircraft that are ready at any time are applied to a small fleet of aircraft. The chapter demonstrates the improvement in computational accuracy that is achieved by reflecting the impact of small numbers of aircraft on availability projections. The approach used is to extend existing finite queuing spares models to including redundancy. Further, the method is used to optimize spares provision with respect to a user specified availability goal. Although the case study for this work is a military combat aircraft application from the Gulf War, the method is applicable to any small system of vehicles or machines where components may be redundant, demand and repairs may be approximated as following an exponential distribution, and limited access to spare parts is the rule.

Chapter 14 presents a condensed synopsis of high velocity maintenance within US Air Force maintenance, repair, and overhaul operations. The Air Force is recognized as a world leader in pushing the limits of technology to overcome constraints to mission success. The weapon system development efforts resulting in the F-117, B-2, F-22 and Global Hawk UAV are clear examples of innovative advanced technologies leading to game-changing effects in the battlespace. Despite the USAF's cutting edge technological supremacy, the maintenance, repair, and overhaul (MRO) of USAF weapon systems has experienced only limited advancement through the application of the full spectrum of IE methods. While USAF MRO operations have made excellent progress through implementing Lean and Air Force Smart Operations for the 21st century (AFSO21) methodologies, these functions and facilities still largely operate as "job shops". In this environment, the craftsman/mechanic accomplishes all work site preparation; gathers parts, tools, and equipment for the day's tasks; acquires all Technical Order (TO) manuals, drawings, and documentation required for each task; and performs a variety of other activities that limit overall productivity.

Chapter 15 presents a comparison of authorized versus assigned aircraft maintenance personnel. The premise of the chapter is that the net effective personnel (NEP) methodology has the potential to be used alone or in conjunction with the Logistics Composite Model to better portray maintenance personnel requirements and capabilities based on experience and skill levels. "Beyond Authorized Versus Assigned: Aircraft Maintenance Personnel Capacity" quantifies the phrase "we need more people" beyond the traditional metric of authorized versus assigned personnel. The chapter is based on work done for a recent Air Force Logistics Management Agency project—C-5 TNMCM Study II. During this project, an extensive, repeatable methodology was developed and utilized to scope an original list of 184 factors down to two potential root causes. These two factors were aligning maintenance capacity with demand, and the logistics departure reliability versus TNMCM paradigm. To address the root cause factor of aligning maintenance capacity with demand, a method of determining available maintenance capacity was needed. To meet this need, a new factor designated as NEP was developed. NEP articulates available maintenance capacity in a more detailed manner that goes beyond the traditional authorized versus assigned viewpoint. The chapter describes how the NEP calculations were developed during the C-5 TNMCM Study II. The NEP calculations were ultimately used in conjunction with historical demand to propose base-level maintenance capacity realignments resulting in projected improvements in the C-5 TNMCM rate.

Chapter 16 presents a discussion of joint and multinational campaign planning using a project/program management approach. An analytical framework is developed which uses a project/program based modeling and management approach to aid in the planning, execution, and control of a joint or multinational campaign. The distinction made here between a project and a program is simply a matter of magnitude, where a program might, in fact, be composed of a number of projects. The framework enables planners to decompose large problems, such as the apportionment of resources to theater commanders, into smaller, more easily managed sub-problems. The sub-problems may similarly be decomposed into even smaller sub-problems. This allows the framework to adapt to virtually any specified degree of resolution. This same hierarchical planning framework is already used in the development and execution of weapons systems planning, development, and acquisition programs. The ability to decompose large problems within this framework hinges on the structure of the problem. When appropriately formulated, the Joint and Multinational Campaign Planning problem demonstrates a block-angular structure most frequently associated with the Dantzig–Wolfe decomposition algorithm. Exploiting this structure not only facilitates problem solution, but also enables a number of analysis options.

Chapter 17 presents a technique for mobilizing Marine Corps officers. The ability to rapidly mobilize the Marine Corps in times of crisis is a cornerstone of US defense strategy. To mobilize rapidly, the marines need an efficient system for assigning officers to mobilization billets. The system designed and built is based on a network optimization algorithm that works in conjunction with carefully designed and scrupulously maintained Marine Corps databases. It takes less than 10 minutes on a 386-based personal computer to complete a mobilization involving 40,000 officers and 27,000 billets and to produce output suitable for generating orders to report via MAILGRAM. Prior to this work, the Marine Corps had a mainframe-based system that took two to four days to complete a mobilization. The new system is not only much faster than the old system, but it also produces significantly better assignments with respect to all measures of effectiveness considered.

Chapter 18 presents the Deployment Scheduling Analysis Tool as an analysis tool for studying military deployment scenarios. The tool has a user interface through which the user selects the deploying units, equipment and their required delivery dates. The user also selects the ports (air, rail, and sea) through which the units will travel to get to the final destination. Finally, the user selects the transportation assets on which the units will travel and assigns those assets to the routes in the deployment network. The system schedules the equipment for deployment by creating individual equipment routings and then repetitively simulating the deployment to determine the deployment closure time and the maximum lateness of any unit. The equipment is scheduled to minimize the maximum lateness of all deploying units. Deployment information is formatted into various graphs and reports. Finally, the user can modify various aspects of the deployment scenario to conduct sensitivity analysis. A valuable option is a heuristic procedure to reassign transportation assets in an attempt to further reduce the deployment closure time and the maximum lateness.

Chapter 19 presents a tool named the Deployment Analysis Network Tool Extended (DANTE) as a network-based analysis tool for studying large-scale deployment scenarios. DANTE was developed originally to support analysis of strategic mobility concepts in an Army Science Board study. DANTE represents a deployment from bases in the US from forward-based locations and from prepositioned stocks as a time-phased network flow. The objective is to minimize the time to close the deploying force on the staging area. The model is optimized using a minimum cost network flow algorithm. A graphic user interface allows deployment parameters to be quickly modified so the impacts of changes on the time to close can be quickly determined. A simple example is given to demonstrate the use of DANTE in analysis of a deployment scenario. DANTE is used by the Military Traffic Management Command–Transportation Engineering Agency.

Chapter 20 presents a methodology for reserve manufacturing capacity for augmenting logistics requirements. The development of logistics and supply chain support for contingency operations is critical for homeland defense and national security. This chapter provides a framework for systematically augmenting production and manufacturing support to meet surging demands for supplies during major military contingencies. A model is presented for constructing alternative allocation strategies along with an example. This approach can be used for developing contingency plans and examining readiness for national defense. American industry plays a major role in national defense by providing the equipment, materials and supplies necessary for military forces to maintain readiness and respond to contingencies that threaten national security. The US has always relied on the private sector to provide increased and surging logistics support for the military during wartime by shifting production plans and priorities to support war efforts. US public and private sectors have always been responsive to support our nation during times of crises and emergencies. Selected industries serve as defense partners through contracts and agreements to augment our national defense effort when called upon to manufacture war materials for contingencies. Most notable contingency manufacturing systems existed during World War II when limited war reserves made it necessary for industries to shift from normal production to produce weapons, ammunition, and supplies for supporting the war effort.

Chapter 21 presents inventory models for contingency operations. Contingencies arise from crises events that require emergency action to assist and protect people from disastrous situations. The logistics functions for contingency operations are complicated due to the prevailing uncertainties and dynamics of demands and mission times. This is particularly crucial for critical inventories that must be managed strategically to ensure mission success. This chapter presents some models for establishing inventories for critical resources such as ammunition, fuel, and medical supplies that are necessary to sustain operations during a contingency and can have a major impact on the mission. The strategy is then to conserve resources to the extent possible. A single period inventory model is developed with time-dependent demand. Results are provided for the case of Poisson distributed demand conditional on random mission time. The public awareness and threat of upheaval from terrorist attacks, natural disasters, and other uncertain crises events has increased dramatically over the past two decades. Unfortunately, we have to posture ourselves to deal with a broad range of contingency events that threaten the safety and security of innocent victims within the United States and throughout the world.

Contingencies arise from crises events that require emergency action to assist and protect people from disastrous situations. They are not restricted to military action, though often military support is provided for disasters occurring within the continental US and abroad. Essentially, all contingency operations require significant logistics functions for providing procurement, storage, transportation, and distribution of supplies and materials for meeting rigorous time and scheduling requirements. While these logistics functions are rather standard for any operational system, they are further complicated in contingency operations due to the prevailing uncertainties and dynamics of demands and mission times. This is particularly crucial for critical inventories that must be managed strategically to ensure mission success. There is no literature that deals with these conditions.

Chapter 22 presents a process of planning ground force operations in the post Cold War era using a systems analysis approach. In the time since this chapter was written the political/military environment has grown much more complex. The construct remains valid, and appropriately updated would provide a basis for supporting analysis. During the Cold War period, the force structure of the US and its NATO allies reflected the requirement to deter and, if necessary, defeat an aggression by the Soviet Union and the Warsaw Pact. It included the capability to deal with "lesser" requirements that had to be addressed occasionally. However, since the end of the Cold War, and Iraq's defeat shortly thereafter, more than 50 military operations have been undertaken to accomplish a series of rather diverse "lesser" missions such as, for example, preventing or ending intra-state conflicts, providing humanitarian assistance and disaster relief, and drug interdiction. This indicates that operations of the "lesser" kind referred to as Stability and Support Operations (SASO) by the US Army have become the rule rather than the exception. Yet the ground forces of the US and most of its allies are still configured for Contingency War Operations (CWO) to deal, as a rule, with military threats of the Soviet type, albeit at considerably reduced force levels. The data published by IISS indicate that the manpower level of US ground forces, including marines, has been reduced by 31% since 1989; that of NATO Europe by 23% (IISS, 2000). SASO-capabilities are generated, with some delay, as the need arises through reconfiguring and retraining combat arms and support units structured and sized for combined arms combat in CWO. On the other hand, there are a few nations, most notably Canada, which have reconfigured, or are about to do so, their ground forces to deal, as a rule, with SASO demands in a responsive manner, however, depriving them of the capability for participating in CWO lest they mount major efforts for rebuilding their forces.

Chapter 23 presents general discussions of supply chain management. There is a great deal of confusion regarding what supply chain management involves. In fact, many people using the name supply chain management treat it as a synonym for logistics or as logistics that includes customers and suppliers. Others view supply chain management as the new name for purchasing or operations, or the combination of purchasing, operations and logistics. However, successful supply chain management requires cross-functional integration within the firm and across the network of firms that comprise the supply chain. The challenge is to determine how to successfully accomplish this integration. In this chapter, supply chain management is defined and the uniqueness of our framework is explained. Descriptions of the transactional and the relationship management views of business process management are provided. The supply chain management processes are described as well as the importance of standard business processes. Then, there is an explanation of how the supply chain management processes can be used to achieve cross-functional and cross-firm integration. There is a description of how customer relationship management and supplier relationship management form the critical supply chain management linkages and how their impact on the financial performance of the organization can be measured. In addition, the partnership model is introduced as means of building high-performance relationships in the supply chain. The ultimate success of the business will depend on management's ability to integrate the company's intricate network of business relationships.

Chapter 24 presents an analytical methodology for hierarchical dynamic decision making. The chapter describes the development of a simulation-based forecasting methodology for hierarchical dynamic decision making (DDM). The forecasting approach accounts for probabilistic factor interactions in decision systems that are characterized by dynamic and probabilistic factors. The methodology uses a hybrid framework of the analytic hierarchy process (AHP), factor analysis (FA), and incorporates a spanning tree (ST) approach. The factor analysis component is used to determine the most representative factors (factor relative weights) influencing the forecast decision. In conventional AHP, the amount of data required for pair-wise comparison matrices could be very cumbersome for large problems. This chapter's methodology requires the input of minimal pair-wise comparison information for a scenario set. Computer simulation is then used to generate additional scenarios based on rudimentary probability information (factor weights) provided by the decision maker. The premise of the approach is that probabilistic interactions exist between the factors that influence the final decision in the AHP model. Potential forecast outcomes are modeled as decision alternatives. The simulation model generates the relative frequency of the selection of each outcome rather than a single final selection. Final forecast decisions are made by the decision maker based on the histogram of the distribution of the simulation results. The methodology facilitates more interactive, data-efficient, and faster decision making under uncertainty.

Chapter 25 presents humans factors in military systems. Human factors engineering, also known as human machine systems engineering is essential to the success and safety of military systems. It revolves around creating systems and artifacts based on the limitations and capabilities (both physical and cognitive) of the human. Military systems tend to be very complex in nature, involving many people, a number of machines, and myriad interactions. The goal of human factors engineering is to integrate the human and the machine in a manner that exploits the best attributes of the two to optimize system performance. The body of knowledge known as human factors is applied in different ways to the design and development of various military systems. This chapter discusses the application of human factors to military communications systems and assistive technologies. This chapter also discusses the application of the tools and techniques of human factors and human–computer interface to the command and control of complex human–machine systems specifically focused on the domain of unmanned aerial vehicles.

Chapter 26 presents Digital Warfighter Modeling for Military Applications. Digital human models have gained significant momentum in recent years for saving money and reducing product cycle time in design, manufacturing, and other areas. Digital warfighters are special digital human models for military applications. In this chapter, a new generation of warfighter is introduced: Santos™, developed at the Virtual Soldier Research Program at The University of Iowa, and applications of Santos in the military. This digital warfighter has features including posture prediction and zone differentiation, cloth modeling, motion prediction, strength and fatigue, physiology aspects, and muscle forces and stress. With these capabilities, this warfighter can be deployed for different military scenarios such as military vehicle design and manufacturing, soldier behavior, and feedback. As a comprehensive human model, Santos provides an example of the kinds of capabilities that can be leveraged with digital human modeling in general. Several examples for digital warfighter applications are demonstrated in the Army, Air Force, Marines, and Navy.

Chapter 27 presents the process of achieving strategic aims in moving toward a process-based military enterprise. Since the end of the Cold War, the US Department of Defense (DoD) has been increasingly called upon to perform everything from toppling standing governments, to nation building and everything in-between. For almost two decades the DoD has struggled to meet these divergent demands with limited resources. Increasingly, the DoD recognizes that a new way of developing, fielding, employing and sustaining capabilities is needed as current requirements are creating resource demands and tasking the DoD beyond its breaking point. The question becomes, how does the DoD chart a course, which when pursued, will yield an organizational construct and associated strategic alignment and execution process which simultaneously meets these very divergent needs at the lowest possible cost to the taxpayer? And more importantly, can it rapidly redirect finite resources to achieve the desired strategic effect? This chapter presents various theories, and introduces the reader to a model which, when properly employed, aligns an enterprise to most optimally meet customer requirements. You will learn a new way for DoD to look at itself—from their customer's perspective, and from that view better conceptualize a new organizational construct and associated strategic planning and execution process which takes the DoD beyond historical stovepipes and simultaneously successfully links strategic aims to measurable tasks.

Chapter 28 presents various philosophies related to building military performance teams. Frederick Taylor (1911), in his work on scientific management, demonstrated the benefit of the performance team. He promoted selection of the best workers for the type of work being performed. He studied the performance processes, and determined a best way to gain the result. Then he gave each worker specific training in that best performance technique. He provided the workers on this team with best tools, best processes, and a manager who was there to do what was necessary or convenient to assure the performance result accomplished by the other team members. As should have been expected, there was no real competition. Taylor's managed work group out-produced the boss-led gang efforts by three to four times. The difference was intelligent management.

Chapter 29 discusses how to initiate performance management systems within the Army. Management engineering is the application of basic industrial engineering in the office environment. It does for management and internal support what traditional industrial engineering has done for efficiency in productive efforts. Application to Army management provides a means to accomplish a larger governmental purpose (performance management) in an area uniquely equipped to do the initial application. This is no add-on management improvement program; it initiates a change that is effective at the level of organizational culture. It discovers an existing performance-management within the Army, and expands it from its present encapsulated area to address Army administration in general.

Chapter 30 presents the Critical Resource Diagramming (CRD) method and its relationship to project work rate analysis. Every real-world project is fueled by the application of resources. Resources are often defined in terms of physical items, tangible property, or intellectual assets. These resources are usually subject to limited availability. Therefore, decisions must be made about resource allocation in order to achieve project goals within the constraints of time, quality, and cost. During the past decades, several researchers have developed heuristic and optimization methods for project activity scheduling, which take resource constraints into consideration. The CRD method is one additional approach to facilitate resource allocation in project scheduling. A CRD is an extension of CPM and it has several advantages which include its simplification in resource tracking and control, better job distribution, improved information to avoid resource conflicts, and better resource leveling. This chapter presents an introduction to CRD as well as analytical formulations designed to optimize the project schedule.

Chapter 31 presents a survey of innovative techniques and practical software tools for addressing military problems. The chapter has six sections. These sections are chosen to illustrate the range and variety of problems confronting the military analyst and the consequent complexity of tools needed to resolve the issues. A secondary goal is to show that with modern turnkey software that these complex tools can be routinely used in an effective way. All of these sections represent actual problems solved by analysts. Some have received recognition by DOD but this will not be indicated in the chapter as the writing focuses on "how to do it". Section 31.1 details a military procurement problem. It is shown that the solution is *based* on a complex dynamic programming model. However implementation requires incorporating recognition of complex regulatory constraints. The savings obtained were verified by audit. Section 31.2 deals with an analysis of the impact of terror on costs of insurance. It is shown that fuzzy arithmetic is an appropriate way to deal with projecting costs and uncertainties. Section 31.3 deals with a problem in maintenance and the readiness derived from maintenance. Here the section is illustrating that choice of metric is sometimes the prerequisite for a successful model and that linear programming cannot be used until the appropriate metric has been found. In this case the utility metric follows the value judgments of actual mechanics and warrant officers who deal with helicopters. Section 31.4 deals with the system design of air dropped sensors and the impact of variability of environment on an element of the system: the battery. It is shown how spread sheet analysis is sufficient to see how variability makes the system vulnerable. This mirrors another classified study on the land warrior system (future Army combat suit) which determined how combat engagement time variability affected battery design. Section 31.5 shows how a turnkey queuing system can estimate the effect of various alternatives in anti ballistic missile defense. In particular this section shows how the meaning of arrival, server, balking, and reneging directly can be translated to ballistic missile defense combat scenarios. For example in a more complex analysis loss of target due to weather can be modeled as reneging and stochastic weather data can be used to enter the parameters for reneging. Section 31.6 demonstrates a technique that is not well known outside military application, that of agent-based models. This is another example of the complexity of the land warrior system and the variety of tools necessary to fully analyze.

Chapter 32 presents computational analysis of countering forgetting through training and deployment. Although worker flexibility has several advantages, it is costly to obtain and maintain given the productivity losses that arise from worker learning and forgetting effects. In this study, factors are reviewed that influence worker forgetting in industrial settings, and the degree to which existing mathematical models conform to observed human forgetting behavior are analyzed. The learn–forget curve model (LFCM) satisfies many characteristics of forgetting. In the context of worker flexibility, LFCM is used to understand the extent to which cross training and deployment become important in helping reduce forgetting effects. Finally, LFCM is enhanced by augmenting it to incorporate the job similarity factor. Sensitivity analysis reveals that the importance of training and deployment policies is reduced as task similarity increases.

Chapter 33 presents the half-life theory of learning curves. The military is very much interested in training troops fast, thoroughly, and effectively. Team training is particularly of importance as a systems approach to enhancing military readiness. Thus, prediction of team performance is of great importance in any military system. In military training systems that are subject to variability and complexity of interfaces, advance prediction of performance is useful for designing training programs for efficient knowledge acquisition and sustainable retention of skills. Learning curves are used extensively in business, science, technology, engineering, and industry to predict performance over time. Most of the early development and applications were in aircraft production particularly for military applications. Military training systems have also taken advantage of learning curve analysis. Over the past several decades, there has been an increasing interest in the behavior of learning–forgetting curves for training and performance analysis. This is because not all knowledge and skills acquired from learning can be retained due to the diminishing effect of forgetting. This chapter introduces the concept of half-life of learning curves as a predictive measure of system performance, which is an intrinsic indicator of the system's resilience. Half-life is the amount of time it takes for a quantity to diminish to half of its original size through natural processes. The common application of half-life is in natural sciences. For example, in physics, the half-life is a measure of the stability of a radioactive substance. In practical terms, the half-life attribute of a substance is the time it takes for one-half of the atoms in an initial magnitude to disintegrate. The longer the half-life of a substance, the more stable it is. Consequently, the more resilient it is. This provides a good analogy for modeling half-life of learning curves with the recognition of diminishing cost (i.e. increasing performance) with respect to the passage of time. This approach adds another perspective to the large body of literature on learning curves. Derivation of the half-life equations of learning curves can reveal more about the properties of the various curves. This chapter presents half-life derivations for some of the classical learning curve models available in the literature.

Chapter 34 presents a measurement scale for assessing organizational readiness for change. Using a systematic item-development framework as a guide (i.e. item development, questionnaire administration, item reduction, scale evaluation, and replication), the chapter discusses the development and evaluation of an instrument that can be used to gauge readiness for organizational change at an individual level. In all, over 900 organizational members from the public- and private-sector participated in the different phases of study with the questionnaire being tested in two separate organizations. The results suggested that readiness for change is a multi-dimensional construct influenced by beliefs among employees that (a) they are capable of implementing a proposed change (i.e. change-specific efficacy); (b) the proposed change is appropriate for the organization (i.e. appropriateness); (c) the leaders are committed to the proposed change (i.e. management support); and (d) the proposed change is beneficial to organizational members (i.e. personal valence).

References

Frederick W. Taylor. 1911. *The Principles of Scientific Management*. Harper Brothers, New York.

IISS. 2000. "Military Balance 1996–1997, 1999–2000." The International Institute for Strategic Studies (IISS), London, UK.

II

Modeling and Optimization

2

Optimizing a Physical Security Configuration Using a Highly Detailed Simulation Model

Tom Marechal
Eindhoven University of Technology

Alice E. Smith
Auburn University

Volkan Ustun
Auburn University

Jeffrey S. Smith
Auburn University

Erjen Lefeber
Eindhoven University of Technology

2.1 Introduction

When a building or a set of buildings needs a physical security system to prevent intrusions, it can be hard to find a configuration that is cost effective but also performs the security task adequately. Especially, when both human guards and security cameras are used it is difficult to find a near-optimal configuration because of the number of potential designs. How many guards are needed and where should each walk? How many cameras are needed, where should each be located and what direction should each point at? These questions are usually answered by using intuition and common sense. The problem when using intuition and common sense is that it is possible to over-secure or under-secure a certain part of the building. Over-securing results in costs that are higher than necessary while under-securing might lead to intruders breaking in successfully, potentially with catastrophic consequences.

In this chapter, a highly detailed simulation model has been used to find a superior security configuration to secure a building. The building has to be protected against intruders who want to steal items. One could, for example, envision a sensitive area where important paperwork is to be protected from being stolen by an external enemy. Or, possibly a weapon or fuel storage depot that has to be secured from enemy intrusions. In this chapter, a stochastic simulation model is coupled with a heuristic optimizer to find a security configuration that is as cost effective as possible while keeping the probability that an intruder breaks in and steals one or more items near zero. The simulation is used to probabilistically evaluate security configurations while the heuristic is used to put forth candidate security configuration designs.

In the next section, a literature review is given. A simulation model of an actual building with guards, security cameras and intruders has been made. The objective functions and the constraints are given in the Section 2.3. In Section 2.4, more details on the simulation are given. The binary problem representation and the newly developed optimization method (hybrid genetic ant optimization, HGAO) are explained in detail in the Section 2.5. In the Section 2.6, the outcome of the HGAO heuristic on two security optimization problems is given. Conclusions are drawn in Section 2.7. Finally, recommendations for further research are given in Section 2.8.

2.2 Literature Review

In this section, literature with respect to the optimization problem and literature with respect to simulation are given. After that a review on heuristics that are strongly related to the heuristic used are provided.

2.2.1 Optimization Problem

Physical security systems to prevent intrusions have been studied in the literature before. However, usually the studies focus on only a small part of the security system, for example, on the working of a single movement sensor or a single laser sensor. Only a very small number of studies focus on more comprehensive security systems. Olario and Nickerson [9] provide an in-depth study on several sensor configurations to protect one single valuable item. In Cavalier et al. [3] several optimization approaches are performed to find a good security sensor configuration for a greatly simplified model. In that study the probability of detection is a decreasing function of distance between intruder and sensor. It is suggested that empirical research methods should be done to investigate in what way distances from different sensors influence detection probabilities. In Jones et al. [5] arcs and nodes are used to evaluate a security system with Markov processes. Success probabilities for different arcs of the intruder path are estimated in advance. The main focus is to find the way at which an intruder has the highest probability to do harm, using low-detailed information. In Jordon et al. [6] performances of complete security systems are tested. In that study the testing is done both by applying low detailed analytical estimates and by applying simulation using human participants. For the latter purpose a real-time connection to a virtual reality model has been used to incorporate real human behavior.

To the best of our knowledge, the only studies on using a highly detailed simulation to evaluate physical security systems are Hamel et al. [4] and Smith et al. [10]. In Hamel et al. [4] an effective sensor network configuration to detect nuclear, biological or chemical attacks is determined. Detailed information such as location and specifications of buildings and different meteorological situations are used. Note that the type of problem in Hamel et al. [4], detecting a nuclear, biological or chemical attack, is quite different from this study. In Hamel et al. [4], mitigation of the threat is an aim over just detection while this chapter will focus on detection. In Smith et al. [10] a simulation using simple agents to represent intruders and guards is introduced. Many realistic features such as doors, windows, locks and security cameras have been implemented in the simulation. The simulation is used to evaluate predefined security configurations.

No literature describes using a highly detailed simulation to *optimize* the security system. In this chapter the simulation of Smith et al. [10] is used as a base. Some additional features have been added and a graphical engine has been built in. Some minor modifications have been made and some additional features have been added to put even more detail into the simulation. Furthermore, the line-of-sight module that evaluates whether an intruder can be seen by a guard or not has been replaced by a faster line-of-sight module, as described in Ustan et al. [11].

2.2.2 Optimization Heuristics

Because of the large search space of the decision variables (camera numbers and locations, guard numbers and paths) and the relatively lengthy simulation needed to evaluate each proposed security configuration, the problem is posed as a binary (combinatorial) problem and solved with a heuristic. The optimization is performed with a new hybrid heuristic, combining aspects of a binary genetic algorithm and of a binary ant colony optimization. Binary genetic algorithms have been used before in various research fields, two of the many examples are Ananda et al. [1] and Yeo and Lu [12]. Binary ant colony optimization has been used before too, for example, Kong and Tian [7] and Nahas and Nourelfath [8]. In Bu et al. [2] a means to turn continuous problems into binary problems has been developed, so continuous problems can be solved as binary problems too, expanding the use of the binary optimization method described in this study.

2.3 Problem Formulation

The simulation, the objective function and the constraints are devised in such a way that several different optimization problems can be solved.

2.3.1 Objective Function

The objective is to minimize the total costs which are all viewed as periodic (annual in this case). Total costs consist of guard salary, yearly camera value deduction, damage to the building caused by an intruder and the value of any stolen items.

$$\begin{array}{c}\text{Total}\\\text{costs}\end{array} = \begin{array}{c}\text{Guard}\\\text{salary}\end{array} + \begin{array}{c}\text{Camera value}\\\text{deduction}\end{array} + \begin{array}{c}\text{Damage to}\\\text{building}\end{array} + \begin{array}{c}\text{Value of}\\\text{stolen items}\end{array} \tag{2.1}$$

Guard salary depends on the number of guards—each additional guard incurs the same additional annual salary. Camera value deduction depends on the number of cameras in the same manner. That is, both guard costs and camera costs are linear. Damage to buildings can consist of the breaking of doors, windows or locks when an intruder attempts to break in. Furthermore, an intruder can also induce damage at other points in the building, for example when s/he is collecting an item to steal (if for instance the item was stored in a locked display case that has to be broken).

Items are only accounted to be stolen if the intruder exits the building with the item. If the intruder gets caught before leaving the building or the area of buildings, the item itself does not contribute to the total costs. Damage that has been done to the building before the moment that an intruder gets caught does contribute to total costs.

The camera value deduction accounts for camera depreciation and possible maintenance costs, as well as power to operate it. In practice the camera value deduction per year is negligible compared to the guard salary; however, there was no reason to leave it out of the objective function. One might at a first glance think that if camera costs are negligible, the numbers of cameras in the optimal case will be extremely large. In Section 2.4, it is explained why this is not the case.

2.3.2 Constraints

While costs are the objective, the security requirement is handled as a constraint. This can be varied to suit the scenario considered. If for example a storage depot for explosives is to be secured, this constraint must be set to a very high probability of intruder detection and apprehension. In this chapter two constraint functions have been created. The minimum required fraction of intruders that get caught by a guard can be set as a constraint (the apprehension constraint). Furthermore, the minimum required fraction of intruders that will be detected at some point can be set (the detection constraint). An intruder is detected either if a guard directly sees the intruder, or if the guard sees the intruder on a surveillance monitor.

$$\text{Fraction of intruders that get caught} \geq h_1 \tag{2.2}$$

$$\text{Fraction of detected intruders} \geq h_2 \tag{2.3}$$

2.3.3 Possible Optimization Problems

In the objective function (Equation 2.1), any of the four factors that contribute to the cost can be turned off if desired. The objective function has been described in this way to enable a variety of security problems to be tackled. If none of the factors are turned off, the objective will be to find the security configuration with lowest yearly costs. In that case some intruders might break into the building, causing damage to doors or windows. It might even be the case that some intruders steal one or more items without being caught. If one would not want intruders to ever succeed in stealing an item, h_1 in Equation 2.2 should be set to Equation 2.1.

One could also choose to turn off the third and fourth factors in the objective function (Equation 2.1) to find the security configuration that is as cheap as possible, ensuring that no more than a fraction of 1-h_1 intruders break in and steal an item. One could also choose to only turn off the fourth factor, for example, if the value of stolen items cannot be expressed in terms of money.

The simulation and optimization are created in such a way that it is also possible to fix the cameras used (or not to use any cameras) and optimize the guards for that camera configuration. It is also possible to consider monitored cameras only without guard paths.

The optimization problems described above are just a few of the possibilities. Because of the fact that the simulation and the optimization are strongly decoupled and because of the fact that many factors can either be turned on or off, the simulation and optimization can be used to solve a variety of optimization problems.

2.4 The Simulation

Simulation is the tool that is used to verify the security designs in this chapter. When a promising design (a security configuration) has been identified, it is fed into the simulation. Based on this input, the simulation produces an output, i.e. the expected total costs, see Equation 2.1. Due to stochastic behavior in the simulation, the simulation is run numerous times for each design.

Although the simulation is very detailed, many assumptions have been made to simulate and run intruder scenarios and to test security configurations. It is hard to devise realistic scenarios that cover the range of possibilities. One should for instance answer the question where intruders would break in, how often they would break in and how they would behave once they were inside. One should also make several assumptions regarding guard behavior. However, to assist with this task the real-time graphical visualization of the simulation can immediately show if anything unrealistic is happening. The currently implemented aspects of the simulation as well as the assumptions made are described in Sections 2.4.1 through 2.4.6.

Section 2.4.1 describes the building that has been used in the simulation. In Section 2.4.2 features of the simulation are given. Section 2.4.3 describes assumptions that have been made in the simulation. In Section 2.4.4 the intruder scenarios are defined. Section 2.4.5 describes the possible guard patrol paths and Section 2.4.6 describes the camera features.

2.4.1 The Building

The purpose of this study is to find a good security configuration for a certain building or set of buildings. The building that was chosen for study is a three floor actual newly constructed building. The building has one elevator and two stairways. All together, there are just over 100 rooms. The building has nine doors to the outside and 99 windows to the outside. A few simplifications have been made: all three floors are considered identical (except for doors to the outside which only exist at the ground floor), doors are assumed to be exactly as thick as walls, and both stairways and the elevator are modeled as rectangular hollow shafts. (Exact building data are not included because of their volume, but are available from the authors.)

2.4.2 Simulation Features

The goal of the simulation is to see how well a certain security configuration behaves for a depicted set of intruder scenarios. In the visualization one can see the intruders and the guards moving around and one can see the cameras. One can also see the areas that guards and cameras can view at each moment in time. Although the graphical representation is a 2-dimensional visualization, the entire model is 3-dimensional. The goal of the real-time graphical representation is to verify that the simulation model works as intended.

Intruders and guards walk a predefined path with a predefined speed at each arc of the path. Behavior of intruders is not determined on the fly, so an intruder will not alter behavior if s/he sees a camera or a guard. "Barriers" are ceilings, floors, walls, doors and windows. Doors and windows are "portals". Doors and windows can be opened, closed or even locked. Intruders can break locks of doors or windows. An agent (guard or intruder) can go to another floor by using the elevator or the stairs.

A line-of-sight algorithm [11] calculates whether a guard can see an intruder or not. In summary, the line-of-sight algorithm first determines a field of vision for each entity or camera, based on its viewing direction, on its maximum viewing distance (a preset parameter) and on the width of its viewing angle (also a preset parameter). After this, the algorithm checks if there are any points of other entities within this field of vision. (Each entity is represented by six points, one on the head, one on each arm, one on each foot and one on the stomach.) If there are any entity points within the field of vision, a check is done to see whether these points are actually visible, or if barriers (such as walls, ceilings or nontransparent doors) obstruct the view. The view of cameras is projected on surveillance monitors located in central bank. Whether the intruder can be seen by the camera is also determined by the line-of-sight algorithm. An intruder who is seen by a camera can only be detected if a guard is actually looking at the surveillance monitors. Only a fixed number of camera views can be projected on the surveillance monitors at the same time, if there are more cameras a switch is made every few time units. Figure 2.1 shows a typical screen shot of the simulation. In this figure, solid black lines are portals and the cameras, guards and intruder are all shown with their field of vision for that moment in time.

2.4.3 Simulation Assumptions

While the simulation used is quite detailed and flexible, certain assumptions regarding the facility and agent behavior must be made. These assumptions are common sense for the most part, however, some security scenarios may need different sets of assumptions:

- Initially all outside doors and windows are locked, some of the inner doors are locked too.
- A guard can go through every door quickly (that is, has a key for locked doors).
- An intruder can go through any door or window, but if the door or window is locked it will take considerably more time.
- Intruders can induce costs at any point in the building, defeating a locked door or window will induce costs, but an intruder can also cause damage at other places.
- A guard does not realize that an intrusion has occurred if s/he detects that the state of a door or window has been changed.
- An intruder walks slower if carrying a heavy item that has been stolen.

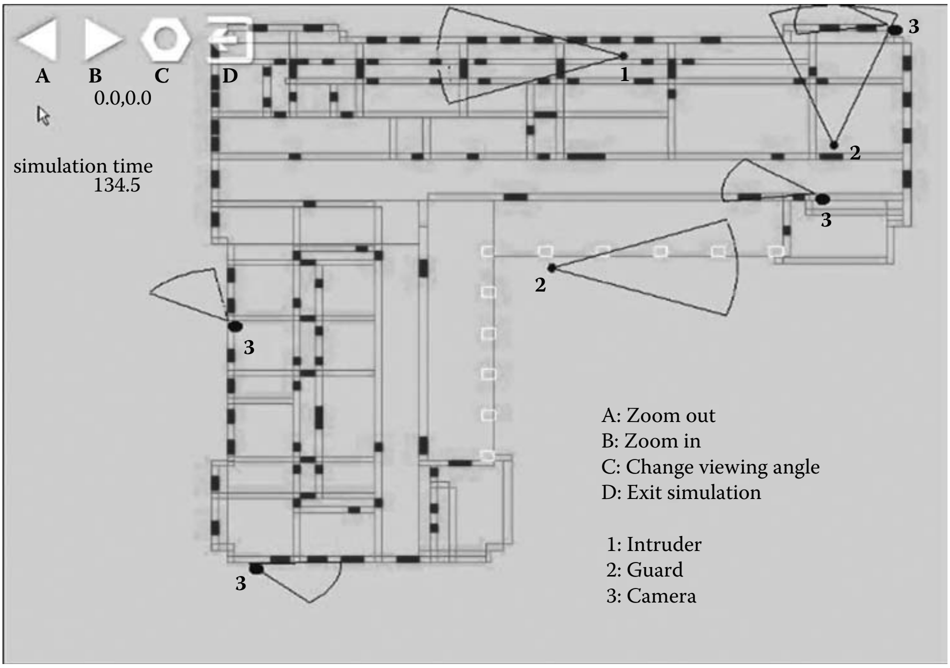


FIGURE 2.1 Screen shot of the 2-dimensional graphic visualization of the simulation.

- There is never more than one intruder at a given time.
- An intruder is detected if a guard sees her/him. A guard sees the intruder with a certain probability if the intruder is in the area the guard is physically capable to see.
- If an intruder is in sight of a camera, the intruder is detected with a certain probability if that camera is projected on the surveillance monitor and a guard is watching the surveillance monitor.
- No attempt has been made to simulate guards that actually chase the intruder. A calculation is done to determine the time an intruder has left after detection before s/he is caught. The amount of time depends on the distance between guards and the intruder, on the guard's running speed and on the floor the guards and intruder are. If the intruder is detected via camera, the guards first need a few seconds to determine where to go to. This delay could be caused by communication between guards on the intruder position for example.
- An item is only considered stolen if the intruder exits the building with it.

2.4.4 Intruder Scenarios

To test the optimization heuristic, certain intruder scenarios, candidate guard patrol paths and candidate camera options and positions have been defined. All of these settings can easily be changed to test for different situations. For this study, 40 intruder scenarios have been defined with the following assumptions:

- The number of times intruders try to break in per year is fixed.
- The largest proportion of the costs caused by an intruder consists of the value of stolen item(s).
- An intruder will only break in via a door or window on the first floor. However, some intruders go to the second or third floor via the stairs or via the elevator after breaking in on the ground floor.
- Intrusion scenarios where intruders go to the second or third floor of the building are less likely to happen than intrusion scenarios where intruders stay at the first floor.
- Some intruders only steal one item, some intruders steel multiple items during one intrusion, some intruders go in and out of the building several times each time stealing one or more items.

- Intruders do not always exit the building at the same place where they entered, but in most cases they do.
- In general, intruders move fast, except when they are carrying a heavy item.

2.4.5 Guard Patrol Paths

Seventeen candidate guard patrol paths have been defined. The following assumptions have been made:

- Some guard paths are outside the building, some are inside the building.
- Guard salary is $31 per hour per guard.
- The building needs to be guarded for 60 hours a week, 52 weeks a year.

2.4.6 Camera Features

Forty-one candidate camera positions have been defined. The following assumptions have been made:

- Some of the cameras are outside the building, others are inside the building.
- Some cameras have predefined scanning movement possibilities, other cameras look in a fixed direction all the time.
- The width of a camera view is 50 degrees.
- The camera costs are $500 per camera per year. These costs are caused by maintenance and camera value deduction. Costs of surveillance monitors, wires and data storage devices are neglected (or considered loaded into the camera cost).

2.5 Optimization Heuristics

Because of the complexity of the objective function evaluation and the large search space size, heuristics are considered for optimization. Heuristics require a solution encoding, a method to evaluate solutions, a method to move to or to identify new candidate solutions and a termination mechanism.

2.5.1 Security Configuration Representation and Evaluation

The problem is described as a binary problem, where each solution consists of two strings with binary digits, one string lists possible camera locations while the second string lists possible guard paths. For each predefined candidate camera location or guard path, there is one digit. If the digit is set to 0 this means that the candidate camera location or guard path is not used, if the digit is set to 1 the candidate camera location or guard path is used in the security configuration. Summing over the camera string yields the number of cameras while the same is true for guards when summing over the guard string.

Each solution is evaluated by the mean of three replications of the simulation. If the mean of these three indicates infeasibility (does not meet the detection/apprehension constraint), the solution is discarded. Otherwise, the solution is considered for inclusion in the population. A more precise simulation set of 30 replications is done for each solution which has a higher fitness than the worst solution in the current population and has a fitness which is not more than 10% lower than the fitness of the best solution in the population. The mean of these 30 replications is used as the evaluation. Thirty replications are chosen as that is an accepted lower bound on "large sample size", where normality of the mean can be assumed. The reason for this two step evaluation is that the simulation takes quite a bit of CPU time, and this method conserves it.*

* For a number of solutions, the 30 replications were also compared with 600 replications. It was ascertained that 30 provided the same results as 600, and therefore for this simulation, was deemed sufficient for optimization.

2.5.2 The Need for an Effective Heuristic

There are 41 candidate camera locations and 17 candidate guard paths resulting in a 58 dimensional binary solution space. The number of cameras and guards used in the security configuration is also determined by the heuristic, therefore there are $2^{58} \approx 2.9 \times 10^{17}$ solutions. Because there is stochastic behavior in the simulation (for example, door or window opening time), each solution has to be tested many times to obtain the distribution of results. Because of this and because the highly detailed simulation is computationally expensive, there is a need for a very effective heuristic.

Two popular heuristics inspired by nature are combined into a new hybrid algorithm, the hybrid genetic ant optimization (HGAO). HGAO combines properties of genetic algorithm (GA) and ant colony optimization (ACO), which for this problem is more effective than either heuristic alone. In general ACO converges faster than GA, because ACO especially depends on local information (pheromone level of each single digit) and GA especially relies on global information (fitness of entire solution). Specifications of the GA and the ACO are given in Sections 2.5.2 and 2.5.3, respectively.

2.5.3 GA

In the genetic algorithm strings can been seen as chromosomes and digits can be seen as genes. For general information on GA, see Ananda et al. [1] for pseudo code on the application of the GA in this chapter, see Figure 2.2. In this chapter, initializing the GA is done as follows: First, random initial solutions are created. This is done by choosing a random number between 0 and 1 for each initial solution, then for each bit in the string, another random number is chosen between 0 and 1. At each bit place, if the random number is larger than string random number, a 1 is assigned for that bit, otherwise a 0 is assigned. After initialization, a genetic algorithm consists of four stages; selection, recombination, mutation and updating.

Selection: The two parents that are used for creating the next child(ren) can be selected in various ways. In this chapter single round tournament selection is used. For each parent two individuals from the population are chosen at random. The fittest of these two individuals will be one of the next parents. In this chapter two parents only create one child.

Recombination: In this chapter equal probability gene cross-over (also called uniform crossover) is used. For each gene, there is a 50% chance that the gene (0 or 1) of the first parent is used and a 50% chance that the gene of the second parent is used.

Mutation: In this chapter gene flipping probability is used. This means that for every gene the probability that the gene flips (either from 0 to 1 or from 1 to 0) is defined. The probability is set to 0.05. Using this probability for gene flipping results in some strings with no mutations and some strings with one or several mutations.

Updating: Usually whenever an entire population has been created and tested, the fittest solutions are selected following a certain strategy and these solutions create a new generation, termed a general

```
Pseudo code GA
Repeat
      Create one random individual
      Evaluate solution
      If feasible
            Add to population
Until (number of individuals = max population size)

Repeat
      Select 2 parents by tournament selection
      Apply genetic operators to generate new individual
      Evaluate solution
      If feasible & fit enough
            Add to population and remove worst individual
Until too many steps without improvement OR max calculation time reached
```

FIGURE 2.2 Pseudo code of the genetic algorithm.

GA. However, because of the computational effort of evaluation, an incremental GA is used here. After being created (by two parents), the child will only be added to the population if its fitness is higher than the fitness of the solution with the lowest fitness in the population. Solutions that are not feasible with respect to the constraints described in Section 2.3 will be discarded immediately.

2.5.4 ACO

In the ant colony optimization algorithm, strings can be seen as paths and each digit can be seen as a dimension. So if a string consists of n digits, the problem is n-dimensional. For each dimension there is a pheromone concentration. This concentration is a real number, the probability that an ant walks into this dimension (that, the probability that the search moves in to this solution), depends on this pheromone concentration. The pheromone levels are initialized by initially setting all values to 0.5 (or, to a random walk). For general information on ACO, see Kong and Tian [7], for pseudo code on the application of the GA in this chapter, see Figure 2.3.

An ant colony algorithm consists of three stages; path picking, pheromone increase and pheromone decrease.

Path picking: In this chapter double probabilistic path picking is used. First the number of digits that will be set to 1 in the next solution will be determined via a probability. The mean of this probability equals the number of digits that were set to 1 in the corresponding string in the previous solution. Once the numbers of digits that will be set to 1 in each string (guard string and camera string) are determined, the actual digits that are set to 1 are chosen in a probabilistic way using the pheromone concentration.

Pheromone increase: For cameras, whenever a camera detects an intruder, the pheromone level for its corresponding digit is raised. For a guard path, the pheromone level is increased when s/he actually catches an intruder. In addition, the pheromone level of a guard increases if the guard detects an intruder on the surveillance monitor. Note that even for solutions that are infeasible with respect to the constraints given in Section 2.3 the pheromone levels can increase.

Pheromone decrease: The ACO in this chapter uses min/max criteria. Whenever a pheromone level exceeds a maximum value, the pheromone levels for all genes on that string will be scaled down by the same fraction in such a way that the highest pheromone level equals the maximum value. Whenever a pheromone level comes below a minimum value, this pheromone level is set to the minimum value. This prevents pathological behavior of the ACO search.

2.5.5 HGAO

In HGAO, each new solution is created either by GA or by ACO (Figure 2.4). The idea behind this is that whenever an ACO solution is created, it will be close to the previous solution in the solution space. If this results in lower fitness, the next solution will probably be created using GA. However, if the ACO

```
Pseudo code ACO
For every bit
      Set initial pheromone concentration
Repeat
      Create one random individual
      Evaluate solution
      Update pheromone levels
Until n random individuals have been created

Repeat
      Create new individual based on pheromone levels
      Evaluate solution
      Update pheromone levels
Until too many steps without improvement OR max calculation time reached
```

FIGURE 2.3 Pseudo code of the ant colony optimization.

```
Pseudo code HGAO
For every bit
      Set initial pheromone levels
Set initial Xp (Xp = probability that next individual is created by ACO
strategy)
Repeat
      Create one random individual
      Evaluate solution
      Update pheromone levels
      If feasible
            Add to population
Until (number of individuals = max population size)
Repeat
      If (random number between 0 and 1 < Xp)
            Create new solution based on pheromone levels
            Evaluate solution
            Update Xp
      Else
            Apply genetic operators to generate new solution
            Evaluate solution

      Update pheromone levels
      If feasible & fit enough
            Add to population and remove worst individual
Until too many steps without improvement OR max calculation time reached
```

FIGURE 2.4 Pseudo code of the hybrid genetic ant optimization.

solution results in a higher fitness, the probability that the next solution will be created by ACO again is high. This provides a balance between the strengths and successes of each heuristic.

Which one of the two heuristics (GA or ACO) is used for creating the next solution depends on the value of variable X_p in the HGAO heuristic. X_p is a real value ranging from X_{min} ($X_{min} \geq 0$) to X_{max} ($X_{max} \leq 1$). The value of X_p directly represents the probability that the next solution will be created using ACO. If a solution is not created by ACO, it is created by GA. X_{min} and X_{max} are static real values, they are both set before the optimization starts. X_p is a real value that changes during the optimization. For this work X_{min} was set to 0.2 and X_{max} was set to 1. This means that at least 20% of the solutions are created using the ACO approach and in the other extreme case all solutions are created using ACO approach.

There are many ways to update X_p. In this work a simple decision strategy is chosen—X_p is only updated after testing an ACO solution. If the ACO solution gave a good result (feasibility and high fitness), the value of X_p will be increased by X_{add}, so the probability that the next solution will be created using ACO will be higher. If the ACO solution gave a bad result (infeasibility or low fitness), the value of X_p will be decreased by $X_{substract}$, so the probability that the next solution will be created using ACO will be lower. X_{add} and $X_{substract}$ are set to the relatively high values 0.4 and 0.2, respectively, so X_p can fluctuate rapidly. The idea behind this decision strategy is that the GA effectively scans the global search space while the ACO part specializes in local optimization.

2.6 Results

The optimization methods and the simulation have been written in such a way that various security optimization problems can be solved. In this section results of two different optimization problems are given. Results of the GA and of the HGAO will be compared. Furthermore an in-depth look at the HGAO heuristic's behavior will be given.

The optimization problem described in Section 2.6.1 is large and complicated. This problem is where all possibilities of the simulation and optimization are tested. The optimization problem in Section 2.6.2 is a smaller one and this problem is chosen to get an overview of the effects of setting constraints and of using alternative numbers of guards. A population of 30 is used by each algorithm for each problem.

2.6.1 Test 1: Unconstrained Full Optimization

The goal of this optimization is to find the lowest total cost security configuration. Both cameras and guards can be used. No constraints are used, that is, no minimum levels of detection or apprehension are set. The reason for optimizing this problem is that it is a large and complicated problem. It would be difficult to find a good security configuration for this problem using only common sense and intuition. Forty intruder scenarios are considered, 41 possible camera locations and 17 possible guard paths. The optimization is terminated when number of intruder runs reaches three million, resulting in testing approximately 14K different security designs and taking about 48 hours of desktop PC computing time. Note that the theoretic search space is $2^{(40+41+17)} = 7.9 \times 10^{28}$.

2.6.1.1 Results

The same problem was optimized using GA and HGAO. Both the GA and the HGAO are run for approximately 14,000 iterations, which takes approximately two days (each algorithm) on a current desktop computer. Recall that the simulation itself is the greatest consumer of computational resources. Figure 2.5 shows a typical run of each algorithm as computational time (and thus, iterations) increase (note that the *x* axis is log scale).

One can see that both algorithms have similar performance. In the beginning they perform exactly the same with respect to the total costs of the best solution so far. This is due to the fact that initial solutions for the GA are created with the same seeding in both methods. The GA has a population size of 30, so it is seeded with 30 individuals.

In the end, HGAO finds a slightly better solution than the GA (see Table 2.1). Interestingly, each optimization ends up in a different optimum. Both find that the only guard that should be used is the guard that watches the surveillance monitors all the time (except when an intruder is detected, in which case the guard leaves the monitors to attempt to apprehend the intruder). Both optimizations find that 27 of the 41 candidate cameras should be used. Cameras specified for the upper floors were chosen to view the elevator or stairwells, as these are the only access points from below. Each method finds a different camera configuration (see Figure 2.6 for specifics). The security configuration of the HGAO results in a 0.2% higher detection probability and a 0.5% higher apprehension probability than the security configuration found by the GA. Due to these probabilities the costs induced by intruders are 3.6% lower for the HGAO configuration. A little more than half of the total costs are security costs (guards and cameras)

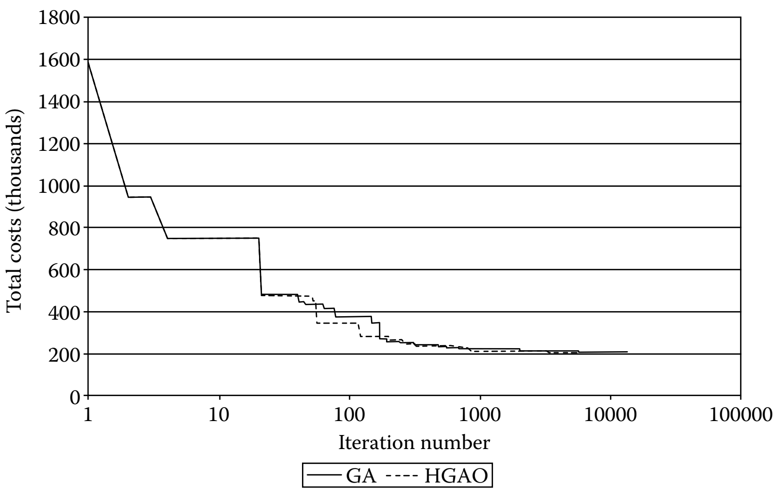


FIGURE 2.5 Total costs of the best solution of each optimization approach as search progresses.

TABLE 2.1 Results of Optimizations with GA and with HGAO

	GA	HGAO
Detection probability	0.967	0.969
Probability of getting caught	0.956	0.961
Number of guards used	1	1
Number of cameras used	27	27
Guard costs	\$96,720-(46.1% of total costs)	\$96,720-(47.6% of total costs)
Camera costs	\$13,500-(6.4% of total costs)	\$13,500-(6.7% of total costs)
Intruder caused costs	\$99,720-(47.5% of total costs)	\$92,572-(45.6% of total costs)
Total costs	\$209,904	\$202,792

while the rest are costs incurred by break ins. In the lowest cost security configuration about 96% of the intruders get caught.

To investigate whether HGAO finds the slightly better solution because of stochastic behavior in the simulations or because the HGAO is a better heuristic for this problem, a more in depth analysis is done in Section 2.6.1.2.

2.6.1.2 In-depth Look at the HGAO Behavior

The idea behind the decision strategy of HGAO is that the ACO part will efficiently search local areas of promising solutions (see Sections 2.5.2 and 2.5.5). This means that in most cases the ACO part should find (near-) optimal solutions for each optimum. If this is the case, the majority of solutions that are good enough to be added to the population should be created by ACO. Furthermore, the best solution that is found at any iteration is most likely to be found by ACO.

However, it turns out that for this problem only 13.3% of the solutions that are fit enough to be added to the simulation are ACO solutions. All other solutions are created by the GA. Of the best solutions so far, only 16.0% were found by the ACO part while the total fraction of solutions created by the ACO was 34.1%. From these statistics, one can infer that an ACO solution is less likely to be a good solution than a GA solution. This outcome contradicts the expected behavior of the HGAO and the fact that HGAO slightly outperformed GA can be attributed to the stochastic behavior of both methods.

2.6.2 Test 2: Constrained Optimization with Guards Only

The goal of the second test is to consider the optimization problem with no cameras. This might be useful in conflict situations where military units move often or in emergency situations (such as just after a natural or manmade disaster) where no camera infrastructure is available. In this case, a constraint on the minimal fraction of intruders that gets caught is set to various values. The trade offs in number of guards, security and total costs and fraction of apprehension will be examined. Up to 17 candidate guards (with their predefined paths) are considered, resulting in a search space of $2^{17} = 131$K. The optimization termination criterion was 300K intruder runs, equivalent to 1597 possible security designs (for each chosen level of security).

Both GA and HGAO found the same security configurations for each optimization, regardless of constraint value. Calculation times were also similar. Therefore no separate results of GA and HGAO are given. Note that in some cases the optimal security configuration is not located on a constraint, but is located in the feasible area slightly away from the constraint. If for instance h_1 is set to 0.95, at least 95% of the intruders must get caught by a guard. Due to the integer number of guards, the outcome might be that if h_1 is set to 0.95 the probability that an intruder gets caught is 0.97. In Figures 2.7 and 2.8 the probabilities of getting caught are not the constraint value, they are the actual probability belonging to the security configuration found. Both figures do not provide results for lower probabilities than 0.899, because if no constraints (or very loose constraints) are used, the best solution (i.e. lowest total costs) has

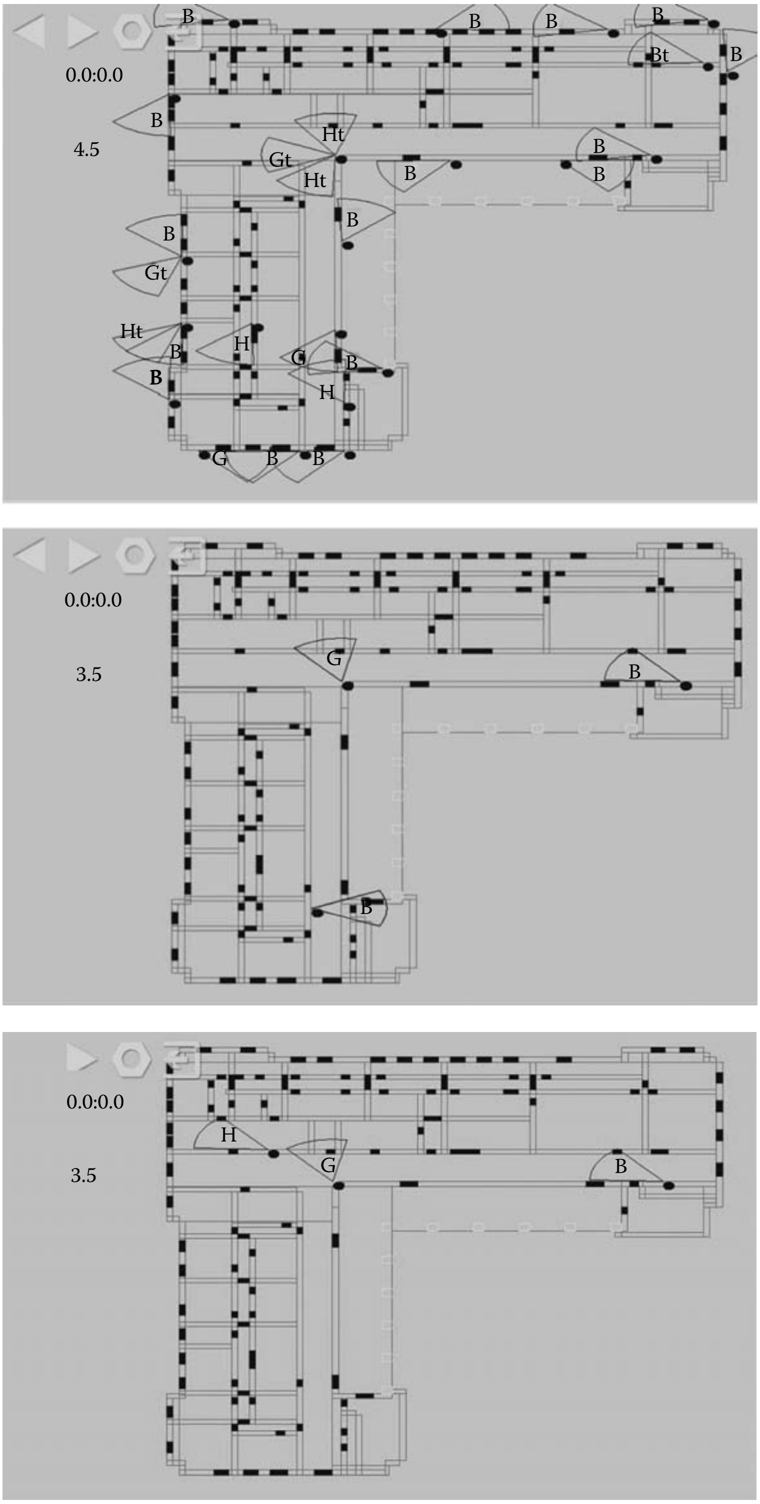


FIGURE 2.6 Optima from HGAO and GA (ground, second and third floors from top to bottom). H: camera only used in HGAO; G: camera only used in GA; B: camera used in both; t: camera can turn, it scans a larger surface.

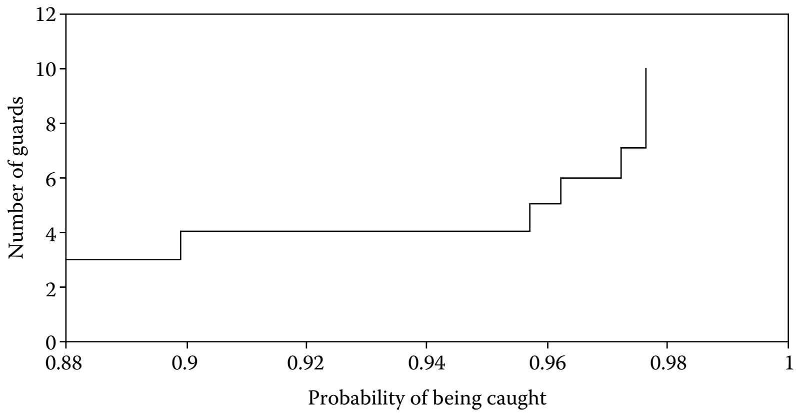


FIGURE 2.7 Number of guards needed versus probability of being caught.

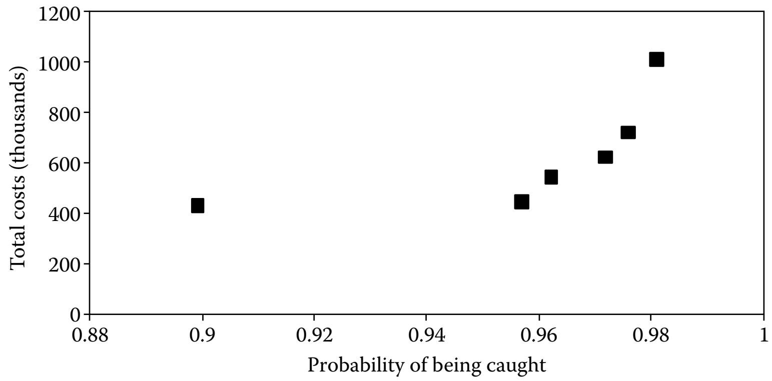


FIGURE 2.8 Total cost versus probability of being caught.

a probability of being caught of 0.899. Furthermore, regardless of constraint level, the final solution of the optimization never has less than three guards.

When more guards are used, more intruders get caught, so the costs caused by intruders decrease. However, when more guards are used, the guard costs increase. From Figures 2.7 and 2.8 one can conclude that for strict constraints, the more guards used, the higher the total costs are. This means that the major part of the total costs must consist of guard salaries. That this is indeed the case can be seen in Table 2.2. This table shows the total costs and the fraction of costs caused by guard salaries.

For the optimal case with respect to total costs, three of the 17 candidate guards are required if no constraints are set. The number of guards needed to obtain a certain probability that an intruder gets caught seems to increase exponentially until it reaches a level of 10 guards (with 98.1% probability of apprehension). After ten guards, no further improvement in the probability of being caught can be made. This is not caused by a flaw in the simulation nor in the optimization heuristic, rather it is due to the fact that the predefined candidate guard patrol paths are not above 98.1% effective. When all 17 possible guards are used, still 98.1% of the intruders get caught. To increase that percentage, more and/or different candidate guard paths would need to be defined. Figures 2.9 and 2.10 show two of the optimal configurations—one with three guards and one with six guards. The three guard solution consists solely of paths around the exterior of the building. With six guards, two more are added to the ground floor, both inside the building, while one is stationed on the second floor.

TABLE 2.2 Costs for Various Probabilities of being Caught for an Intruder Found by HGAO

Probability of being Caught	Guard Salary Costs	Intruder Caused Costs*	Total Costs
0.899 (optimal with respect to total costs)	$290,160 (67.5% of total costs)	$139,566 (32.5% of total costs)	$429,726
0.957	$386,880 (86.8% of total costs)	$58,838 (13.2% of total costs)	$445,718
0.962	$483,600 (89.2% of total costs)	$58,525 (10.8% of total costs)	$542,125
0.972	$580,320 (92.1% of total costs)	$49,806 (7.9% of total costs)	$630,126
0.976	$677,040 (93.0% of total costs)	$51,012 (7.0% of total costs)	$728,052
0.981	$967,200 (95.6% of total costs)	$44,825 (4.4% of total costs)	$1,012,025

* Note that the costs caused by the intruder for a probability of being caught of 0.976 are slightly higher than for a probability of 0.972. This is caused by stochastic behavior of the simulation.

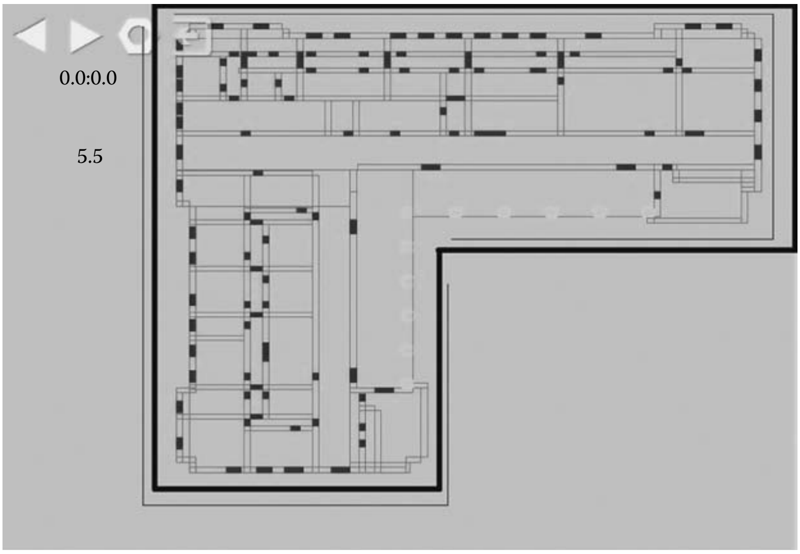


FIGURE 2.9 Best solution with three guards (probability of apprehension = 89.9%). Thick line: guard walks in circles. Other lines: guards walk over this line back and forth.

2.7 Discussion

This chapter described the use of heuristic optimization coupled with detailed stochastic simulation to design physical security systems. Because of the computational effort needed for the simulation, the optimization routine needs to be efficient as it searches a very large combinatorial space. The heuristics described have the advantage that they can be terminated at any point and still return good solutions, and in fact, almost always experience diminishing returns over search time with respect to the objective function (see Figure 2.5 for example). Furthermore, the objectives and constraints are quite flexible in this method so that various problems could be posed and solved using the same simulation and the same optimization framework. Trade offs between alternative security levels and the investment needed can be ascertained.

There are drawbacks to the approach. One is that candidate guard patrol paths, camera locations and details, and possible intruder scenarios have to be defined in advance. Defining them can be time consuming as well as conceptually challenging. The optimization only uses these predefined data, therefore the security design consists of elements among the prespecified ones. A second drawback is the optimization itself. Heuristics are not guaranteed to find optima, and it is even difficult to know how to close to optimal the results are. The heuristics herein discard infeasible solutions, that is those which do not

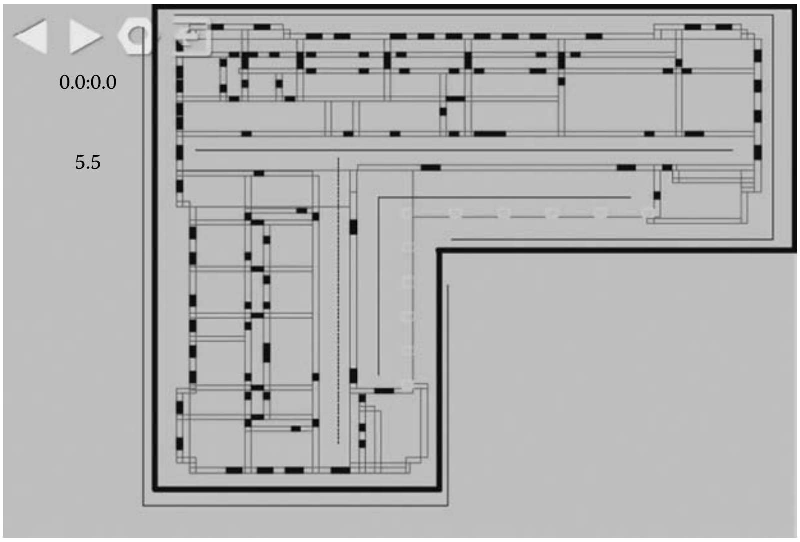


FIGURE 2.10 Best solution with six guards (probability of apprehension=97.2%). Thick line: guard walks in circles. Other lines: guards walk over this line back and forth. Dotted line: guard walks on second floor. On all other lines the guard walks on the ground floor (or outside).

meet the security constraint. It may be advantageous to include these in the population as a source of valuable information for the search mechanism. A final drawback is the simulation model. Adding more detail to it can make things even more realistic. Currently, intruder behavior is independent of security (that is, guards and cameras). This is clearly not the actual case and the simulation would be improved by adding the facility for guards, cameras and intruders to interact. This might be done by creating a network of arcs and nodes and let intruders and guards decide which arc to take at each node on the fly, based on one or several decision variables. Guards should be able to notice certain state changes (for example opening or closing doors) that have occurred, which would indicate an intrusion without seeing the intruder. Guards could then act upon this observation.

This work could be straightforwardly extended by including a wider range of intruder scenarios and studying the scenarios with alternative security set ups in a rigorous manner. A design of experiments might be devised to systematically analyze outcomes and identify significant main effects and interactions. Another straightforward extension is to alter some of the assumptions regarding intruder and guard behavior. Examples are to relax the one intruder at a time assumption or to model behavior in which a guard chases an intruder. These would add realism to the system but would take additional simulation coding and verification, as well as additional computational resources during scenario testing.

However, as a first step in combining detailed simulations with optimization for design of secure facilities, this work offers value and promise. The method is not designed to be "black box"; rather it is designed to supplement human decision making to effect rigorous and defensible decisions. Certainly, when designing for securing, it is vital to know to the best degree possible the ramifications of alternative designs, and to identify all superior designs. This method accomplishes these goals effectively, and should thereby be a valuable tool for security decisions.

References

1. Ananda Rao, M., Srinivas, J., Murthy, B.S.N. 2004. Damage detection in vibrating bodies using genetic algorithms. *Computers and Structures*, 82, 963–968.

2. Bu, T-M., Yu, S-N., Guan, H-W. 2004. Binary-coding-based ant colony optimization and its convergence. *Journal of Computer Science and Technology English Language Edition,* 19(4), 472–479.
3. Cavalier, T.M., Conner, W.A., Del Castillo, E., Brown, S.I. 2007. A heuristic algorithm for minimax sensor location in the plane. *European Journal of Operational Research,* 183, 42–55.
4. Hamel, D., Chwastek, M., Farouk, B., Dandekar, K., Kam, M. 2006. A computational fluid dynamics approach for optimization of a sensor network. *IEEE International Workshop on Measurement Systems for Homeland Security, Contraband Detection and Personal Safety IEEE,* Cat No. 06EX1330C, 38–42.
5. Jones, D.A., Turnquist, M.A., Nozick, L.K. 2005. Simulation of imperfect information in vulnerability modeling for infrastructure facilities. *Proceedings of the 2005 Winter Simulation Conference,* Orlando, Florida, eds Kuhl, M.E., Steiger, N.M., Armstrong, F.B., Joines, J.A., The Society for Computer Simulation International (SCS), San Diego, CA, pp. 965–971.
6. Jordan, S.E., Snell, M.K., Madsen, M.M., Smith, J.S., Peters, B.A. 1998. Discrete-event simulation for the design and evaluation of physical protection systems. *Winter Simulation Conference Proceedings,* Washington, DC, The Society for Computer Simulation International (SCS), San Diego, CA, pp. 899–907.
7. Kong, M., Tian, P. 2006. A new ant colony optimization applied for the multidimensional knapsack problem. *Lecture Notes in Computer Science,* 4247, pp. 142–149.
8. Nahas, N., Nourelfath, M. 2005. Ant system for reliability optimization of a series system with multiple-choice and budget constraints. *Reliability Engineering and System Safety,* 87, 1–12.
9. Olario, S., Nickerson, J.V. 2005. Protecting with sensor networks: perimeters and axes. *Military Communications Conference (MILCOM) 2005, Atlantic City, New Jersey,* 3, IEEE, New York, pp. 1780–1786.
10. Smith, J.S., Peters, B.A., Jordan, S.E., Snell, M.K. 1999. Distributed real-time simulation for intruder detection system analysis. *Proceedings of the 1999 Winter Simulation Conference,* Phoenix, Arizona eds Farrington, P.A., Nembhard, H.B., Sturrock, D.T., Evans, G.W., The Society for Computer Simulation International (SCS), San Diego, CA, pp. 1168–1173.
11. Ustun, V., Yapicioglu, H., Gupta, S., Ramesh, A., Smith, J.S. 2005. A conceptual architecture for static features in physical security simulation. *Proceedings of the 2005 Winter Simulation Conference,* Orlando, Florida, eds. Kuhl, M.E., Steiger, N.M., Armstrong, F.B., Joines, J.A., The Society for Computer Simulation International (SCS), San Diego, CA, pp. 958–964.
12. Yeo, B.-K., Lu, Y. 1999. Array failure correction with a genetic algorithm. *IEEE Transactions on Antennas and Propagation*, 47(5), 823–828.

3

A Time-Window Constrained Scheduling of Target Illuminators in Naval Battle-group Antiair Warfare

Youngho Lee
Korea University

Hanif D. Sherali
Virginia Polytechnic Institute and State University

3.1 Introduction

Consider a naval battle group defending itself against an air attack, equipped with surface-to-air missiles (SAMs), sensors, and target illuminators (homing devices that assist in striking the targets) (see Figure 3.1) [22]. The sensors track the approaching targets (airplanes, cruise missiles), and provide in-flight guidance for the SAMs. Each target must be illuminated during some last few seconds of the SAM trajectory in order for it to home in on the target. Because of the common theater that these battle-group ships operate within, this poses a cooperative engagement problem that calls for a coordinated use of antiair defense resources among the ships. In addition, the number of illuminators that are jointly possessed by the battle group is very small compared to the number of SAMs, and compared to the number of potential incoming targets. Thus, the task of allocating the targets to the illuminators emerges as a critical problem. Additionally, each target must be illuminated during a specified time interval, called an *engageability duration*, that depends on the location of each illuminator. However, the illuminators may be unavailable during certain specified durations due to previously scheduled commitments.

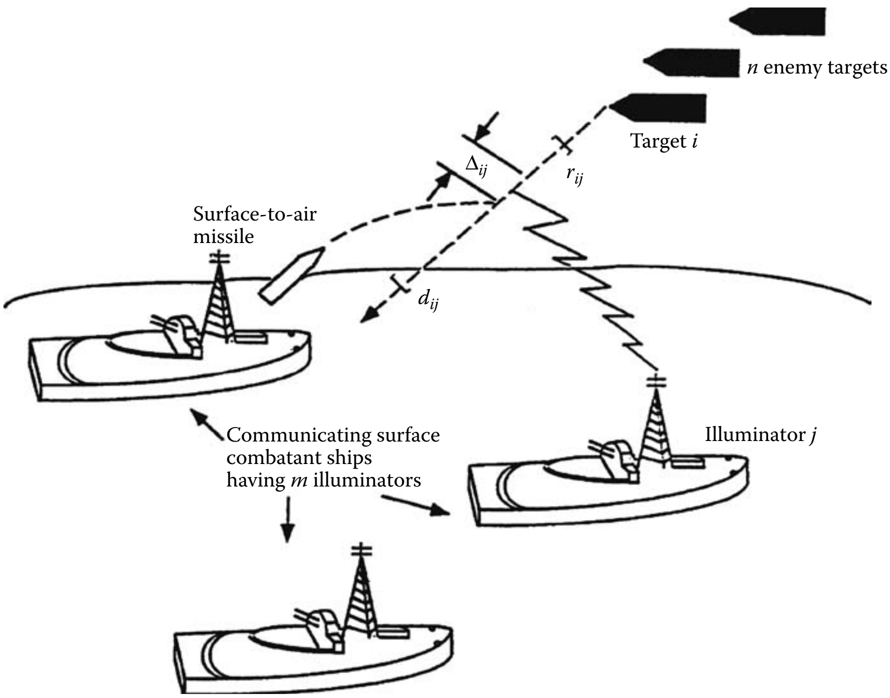


FIGURE 3.1 Illustration of the cooperative battle-group antiwar defense problem.

The cooperative battle-group antiair defense problem can then be stated as follows: *Given a collection of targets approaching on predicted trajectories, and given the engageability durations and the illumination times that are specific to each target and to the illuminators on each individual ship, allocate the targets among the illuminators and schedule them for illumination so as to minimize some appropriate merit or reward function, while satisfying the illuminator availability constraints, as well as the target engageability duration constraints.* In this context, note that the illumination time includes the duration required by the weapons guidance system to lock-in. The subsequent scheduling of SAMs is treated as a separate problem (not addressed herein), which is simpler due to the adequate number of available missiles. Also, this battle group problem is repeatedly solved on a rolling horizon basis to accommodate the dynamic addition of new incoming targets within the system.

As described next, this problem has been conceptualized in Lee and Sherali [14] as a parallel (unrelated) machine scheduling problem of minimizing the total weighted flow time, subject to time-window job availability and machine downtime side-constraints. Suppose that there are n jobs, indexed $i \in N=\{1, \ldots, n\}$ and that there are m machines, indexed $j \in M=\{1, \ldots, m\}$. Each job i has a specified availability duration $[r_{i,j}, d_{i,j}]$, comprised of a release time $r_{i,j}$ and a due-date (deadline) $d_{i,j}$, which is also dependent on machine j. Hence, if job i is assigned to machine j, then it must be processed for a given duration $\Delta_{i,j}$ belonging to the corresponding time-window interval, without interruption or job splitting. In addition, there may be certain specified blocks of duration for which each machine might be unavailable due to scheduled machine maintenance, shift changes, and scheduled breaks, for example. Accordingly, a feasible schedule is comprised of an assignment of jobs to machines and their scheduling thereon, such that each job is processed within its own as well as the corresponding machine's availability time intervals, and such that the processing durations on each machine are all nonoverlapping. Among such feasible schedules, we are interested in finding a schedule that minimizes the total weighted flow time (see Section 3.6 for a discussion of the choice of the objective functions, as motivated by the Naval battle group problem). We denote this *unrelated machine scheduling problem* by *Problem USP.*

From the perspective of computational complexity, determining whether or not even a feasible schedule to Problem USP exists is NP-complete in the strong sense [6,26]. Therefore, theoretically, we may not expect a polynomial-time or even a pseudopolynomial-time algorithm for Problem USP, unless P = NP. Moreover, the problem remains strongly NP-complete even for the single-machine case, and even if only two different integer values exist for the release times and the deadlines [26]. However, as we shall see, the problem does possess some very special structures that permit the development of effective solution procedures in practice.

The remainder of this chapter is organized as follows. In the next section, we develop a zero-one integer programming formulation for this problem. In Section 3.3, we present strong valid inequalities for tightening the initial formulation USP, that can be used in the context of a branch-and-cut procedure [19]. For the purpose of real-time processing, we address the development of several effective heuristic algorithms in Section 3.4 that exploit the special structures inherent in the proposed formulation of the problem. In Section 3.5, we address the development of a branch-and-bound algorithm to find an exact or an approximate solution to Problem USP. This branch-and-bound algorithm utilizes a Lagrangian dual formulation for obtaining a tight lower bound, and employs a heuristic based on Lagrangian relaxation for deriving a good quality upper bound for Problem USP. In Section 3.6, we present an application of the branch-and-bound algorithm for solving the combinatorial naval defense problem described above. We also present computational results using suitable test data. Section 3.7 concludes this chapter.

3.2 Zero-one Integer Programming Formulation

To construct a suitable mathematical formulation of Problem USP, let us assume that the processing times $\Delta_{i,j}$, and the job availability time-window interval endpoints $r_{i,j}$ and $d_{i,j}$, as well as the endpoints of each unavailability interval duration on each of the machines $j \in M$, are all nonnegative integer-valued data. Under this assumption, letting $T \equiv \max_{i,j} \{d_{i,j}\}$ be the problem time horizon given by the latest due-date, the interval $[0, T]$ may be decomposed into slots of unit duration, distributed over all the machines. Let $k \in K$ index all the available (or unassigned) slots, numbered sequentially over all the machines. Let Q be the set of separate pieces of contiguous machine availability intervals, indexed by $q \in Q$, over all the machines and over the time horizon $[0, T]$. Define the sets $T_q = \{k \in K$: slot k is part of the interval $q\}$, for each $q \in Q$. Note that each $T_q \subseteq K$ corresponds to slots that lie within the time interval q, for $q \in Q$.

To formulate Problem USP, let τ_k be the upper interval endpoint of slot k, and let μ_k be the index of the machine that corresponds to slot k, for each $k \in K$. For each job $i \in N$, define a set

$$S_i = \{k \in K: [\tau_k - \Delta_{i,\mu_k}, \tau_k] \subseteq [r_{i,\mu_k}, d_{i,\mu_k}] \text{ and } (\tau_k - \Delta_{i,\mu_k}, \tau_k) \cap [\text{unavailability intervals for machine } \mu_k] = \emptyset\}.$$

Note that the set S_i denotes the slots $k \in K$ for which job i can be feasibly scheduled (independently) to complete by time τ_k. Accordingly, define the binary decision variables $x_{ik} = 1$ if job i is scheduled to complete at the end of slot k, and $x_{ik} = 0$, otherwise. Observe that slot k corresponds to some machine $\mu_k \in M$ and, hence, if $x_{ik} = 1$, then job i has implicitly been assigned to machine μ_k. The cost in the objective function for this assignment is $c_{ik} = w_{ik}\tau_k$ for $k \in S_i$, $i \in N$, where w_{ik}, $k \in S_i$, $i \in N$, are some positive weights (see Section 3.6 for a specific choice of such weight functions). Furthermore, in order to ensure that the schedules of the n jobs do not overlap on any of the machines, let us define for each slot $k \in K$, the set

$$J_k = \{(i,\rho): i \in N, \rho \in S_i, \mu_\rho = \mu_k, \quad \text{and} \quad [\tau - 1, \tau_k] \subseteq [\tau_\rho - \Delta_{i,\rho}, \tau_\rho].$$

Note that for each $k \in K$, J_k is the set of combinations (i, ρ) such that slot k will be occupied if $x_{i\rho} = 1$. Then, Problem USP can be formulated as follows.

USP:

$$\text{Minimize} \sum_{i \in N} \sum_{k \in S_i} c_{ik} x_{ik}$$

$$\text{subject to} \sum_{k \in S_i} x_{ik} = 1, \qquad \text{for } i \in N,$$

$$\sum_{(i,\rho) \in J_k} x_{i\rho} \le 1, \qquad \text{for } k \in K,$$

$$x_{ik} \in \{0,1\}, \qquad \text{for } k \in S_i,\ i \in N.$$

Problem USP includes as special cases several multimachine scheduling problems having release times, due-dates, and time-window constraints [24]. For example, it includes the single machine scheduling problem with release times and due-dates [10,15,20], the multiple machine scheduling problem without time-window constraints [21], problems dealing with the scheduling of unit-time processing jobs having release times and due-dates [25,26], and two-processor scheduling problems with start-times and due-dates [7]. Akker et al. [1] developed a column generation approach for solving machine scheduling problem with the time-indexed formulation that is similar to our proposed model USP. Another time-indexed formulation was proposed for solving a project scheduling problem with labor constraints, along with LP based heuristics [3]. Wolsey [28] has compiled alternative modeling approaches for production planning and scheduling problems including the time-indexed formulation.

Sherali et al. [22] describe a naval defense application as a special instance of Problem USP, and our computational experience in Section 3.6 is based on this application. The proposed formulation for this problem is shown to theoretically, as well as empirically, dominate other alternative models. They design primal and dual heuristics based on a Lagrangian dual approach. In this chapter, we present some classes of tight valid inequalities by exploiting the inherent structural properties of Problem USP. Furthermore, we develop an effective branch-and-bound algorithm that can be used to provide solutions that are guaranteed to lie within any desired percentage of optimality. We mention here that Sousa and Wolsey [27] have also independently presented a similar formulation for a single machine case with only job availability time-window constraints, and have constructed some different classes of valid inequalities along with a linear programming based branch-and-bound algorithm for this special case. Favorable computational results have been reported. Elshafei et al. [4] presented a similar target tracking problem in the context of radar pulse interleaving for multitarget tracking and have developed some effective Lagrangian dual based heuristics. Maskery and Krishnamurthy [16] investigated the problem of deploying countermeasures against antiship missiles from a game-theoretic perspective in which multiple ships coordinate to defend against a known missile threat. In particular, they presented numerical results illustrate the trade-off between communication and performance.

3.3 Strong Valid Inequalities

The initial formulation of Problem USP can be further tightened by introducing some strong valid inequalities. These valid inequalities can be used in the framework of a branch-and-cut scheme by solving a readily available separation problem, or by implementing an ad-hoc procedure for finding a valid inequality that cuts off a fractional optimal solution obtained for the linear programming relaxation of Problem USP, denoted by $\overline{\text{USP}}$. Below, we present an equivalent set packing reformulation of Problem USP, and discuss the derivation of strong valid inequalities via certain knapsack and generalized assignment relaxations of this problem, as well as from cliques of the intersection graph associated with the set packing constraints themselves.

We also illustrate via examples how such inequalities can be generated to tighten the linear programming relaxation of USP. However, for the particular application studied in Section 3.6, we found it sufficient to exploit the special structure identified in Section 3.1, without resorting to additional constraint generation procedures, in order to effectively solve this problem.

3.3.1 An Equivalent Set Packing Reformulation USPP of Problem USP

We can transform Problem USP into an equivalent set packing problem with all less-than-equal-to type constraints, and having a modified objective coefficient vector $\bar{c}_{ik}=\theta-c_{ik}$, for a sufficiently large number θ. This can be done by writing its objective function as maximize $-\sum_{i\in N}\sum_{k\in Si} c_{ik}x_{ik}$ and including the (penalty) term $\theta\sum_{i\in N}(\sum_{k\in Si} x_{ik}-1)$ in this objective function. Adopting this transformation, we obtain the following equivalent reformulation of Problem USP, denoted by USPP, where we have further decomposed the variables over the available blocks $q\in Q$ in order to expose certain special structures.

USPP:

$$\text{Maximize} \sum_{q\in Q}\sum_{i\in N}\sum_{k\in(T_q\cap S_i)} \bar{c}_{ik}x_{ik}-n\theta$$

$$\text{subject to} \sum_{k\in S_i} x_{ik}\le 1, \qquad \text{for } i\in N, \tag{3.1}$$

$$\left[\sum_{(i,\rho)\in J_k} x_{i\rho}\le 1 \text{ for } k\in T_q\right], \qquad \text{for } q\in Q, \tag{3.2}$$

$$x_{ik}\in\{0,1\}, \qquad \text{for } k\in S_i,\ i\in N. \tag{3.3}$$

Remark 3.1. Special structures: Observe that Models USP and USPP enjoy certain inherent special structures. These structures can be identified as follows: (i) a generalized upper bounding (GUB) structure of constraints (Equation 3.1); (ii) a block-diagonal (angular) structure of the set-packing constraints (Equation 3.2); and also (iii) an interval matrix structure of the set-packing constraints (Equation 3.2). (A zero-one matrix is called an *interval matrix* if in each column, the ones appear consecutively. In particular, since the column of each variable x_{ik} in Model USPP has $\Delta_{i,\mu k}$ consecutive ones in the set-packing constraints (Equation 3.2), when the slots are numbered consecutively machine by machine, the block-diagonal set-packing constraints (Equation 3.2) possess such an interval matrix structure.) It is well-known that this interval matrix can be equivalently transformed into a network (see Nemhauser and Wolsey [17]). This can be done by writing constraints (Equation 3.2) as equalities after introducing slack variables, and then by premultiplying these constraints with a nonsingular lower triangular matrix that essentially subtracts row i from row $(i+1)$ for each $i=1, \ldots, |K|-1$. By adding a redundant row that equals the negative sum of the resulting rows, we see that each column of variables (including slack columns) has exactly two nonzero entries, which are $+1$ and -1, and that the right-hand-side has $+1$ in row 1, -1 in the new redundant row $|K|+1$, and zeros elsewhere. Hence, recognizing the inherent network structure produced under such a transformation, Problem USP can be interpreted as a shortest path problem from node 1 to node $|K|+1$ on an acyclic directed graph (digraph), but with an additional restriction due to the GUB constraints (Equation 3.1). Treating Equation 3.1 as equalities as a consequence of the objective functions, this restriction asserts that the path must include exactly one arc from each of the disjoint arc sets $E_1, E_2, \ldots, E_n,$ where each E_i, $i\in N$, contains arcs corresponding to the variables x_{ik}, for $k\in S_i$. These special structures play an important role not only in developing classes of valid inequalities, but also in constructing effective solution procedures, as we shall see in the sequel.

3.3.2 A Class of Valid Inequalities from a Generalized Assignment Relaxation of USPP

We now consider a class of valid inequalities for Model USPP by exploiting the GUB structured constraints (Equation 3.1) and the interval matrix structured constraints (Equation 3.2). Denote as CHUSPP the convex hull of zero-one vertices of the polytope associated with USPP, that is,

$$\text{CHUSPP} = \text{conv}\{x \in \boldsymbol{R}^{n_0} : x \text{ satisfies Equations 3.1 through 3.3}\},$$

where

$$n_0 = \sum_{i \in N} |S_i|.$$

In order to generate a class of valid inequalities for CHUSPP, consider the following surrogates of the constraints (Equation 3.2), corresponding to each machine availability interval $q \in Q$,

$$\sum_{i \in N} \sum_{k \in T_q \cap S_i} \Delta_{i,\mu_k} x_{ik} \le |T_q|, \text{ for } q \in Q. \quad (3.4)$$

Inequalities (Equation 3.4) assert that the total job occupancy time over the interval q is no more than the total available time $|T_q|$. Note that Δ_{i,μ_k} for each $k \in T_q \cap S_i$ in the time block represents the processing time for job $i \in N$ on the particular machine μ_k that corresponds to the slots $k \in T_q$. Consider the relaxed, *GUB constrained, generalized knapsack polytope* (USPKP) associated with Equations 3.1 through 3.4), that is,

$$\begin{aligned} \text{USPKP} = \text{conv}\{x \in \boldsymbol{R}^{n_0} : & \sum_{k \in S_i} x_{ik} \le 1, && \text{for } i \in N, \\ & \sum_{i \in N} \sum_{k \in T_q \cap S_i} \Delta_{i,\mu_k} x_{ik} \le |T_q|, && \text{for } q \in Q, \\ & x_{ik} \in \{0,1\}, && \text{for } k \in S_i, i \in N\}. \end{aligned}$$

Since a facet of USPKP is also a valid inequality for both problems CHUSPP and USP, we investigate the facial structure of USPKP. It can be easily verified that all nonnegative constraints $x_{ik} \ge 0$ are facets of USPKP. These constraints are referred to as *trivial facets* for USPKP. In addition, it is shown by Hammer et al. [11] that, since the constraint matrix coefficients of USPKP are nonnegative and all constraints are less-than-or-equal-to type inequalities, all nontrivial facets are of the form $\sum_{i \in N} \sum_{k \in S_i} \pi_{ik} x_{ik} \le \pi_0$, having $\pi_{ik} \ge 0$ and $\pi_0 \ge 0$.

We can further restrict our investigation of the facial structure of USPKP as follows. For each $q \in Q$, let $y_{iq} = \sum_{k \in T_q \cap S_i} x_{ik}$ for all $i \in N$ such that $T_q \cap S_i \ne \varnothing$. For each $q \in Q$, denote by *j(q)* the common value of μ_k for all $k \in T_q$. Consider the convex hull of the following *generalized assignment polytope,* denoted by USPGAP, where $n_1 = \sum_{q \in Q} |i \in N : T_q \cap S_i \ne \varnothing|$, and where undefined variables are assumed to be zero.

$$\begin{aligned} \text{USPGAP} = \text{conv}\{y \in R^{n_1} : & \sum_{q \in Q} y_{iq} \le 1, && \text{for } i \in N, \\ & \sum_{i \in N} \Delta_{i,j(q)} y_{iq} \le |T_q|, && \text{for } q \in Q, \\ & y_{iq} \in \{0,1\}, && \text{for } i \in N, q \in Q\}. \end{aligned}$$

Remark 3.2. Gottlieb and Rao [8,9] describe some classes of facets for the generalized assignment problem GAP. All these facets can be implemented as valid inequalities for Model USP, because if $\sum_{i\in N}\sum_{q\in Q}\pi_{iq}y_{iq}\leq\pi_0$ with $\pi_0\geq 0$ is a facet of USPGAP, then

$$\sum_{i\in N}\sum_{q\in Q}\sum_{k\in T_q\cap S_i}\pi_{iq}x_{ik}\leq\pi_0 \tag{3.5}$$

is a facet of USPKP, which consequently is a valid inequality for Problem USP. This result follows by noting that since $\sum_{i\in N}\sum_{q\in Q}\pi_{iq}y_{iq}\leq\pi_0$ is a facet of USPGAP, and since $y_{iq}=\sum_{k\in Tq\cap Si}x_{ik}$, for all (i, q), is feasible to USPGAP for any feasible solution x to USP, the inequality (Equation 3.5) is valid for USPKP, and consequently, is valid for Problem USP. The facetial property is a consequence of the converse, namely that if y is feasible to the constraints defining USPGAP, then for each variable y_{iq} that is equal to 1, by putting some corresponding variable $x_{ik}=1$ for $k\in T_q\cap S_i$, and letting the remaining x-variables be zeros, we would construct a solution $x\in$USPKP. This follows from the fact that, for any i and q, the variables x_{ik} for all $k\in T_q\cap S_i$ have identical columns in USPKP. Hence, a formal proof of the foregoing assertion can be readily constructed [13].

In particular, the following proposition provides a separation procedure in order to obtain a class of valid inequalities for Problem USP from each individual knapsack constraint of USPGAP.

Proposition 3.1. All nontrivial facets associated with the individual knapsack constraints of USPGAP are also facets of USPGAP, which consequently correspond to facets of USPKP via Equation 3.5.

Proof: The proof follows by Proposition 5.10 in Gottlieb and Rao [8] and by Remark 3.2.

Remark 3.3. We can solve a standard separation problem (see Nemhauser and Wolsey [17]) in order to (possibly) obtain a minimal cover inequality from a knapsack constraint of USPGAP that cuts off a given fractional solution to USP. A sequential lifting procedure can polynomially strengthen such a minimal cover inequality to a facet of the knapsack polytope [29], and by Remark 3.2, such a constraint would then correspond to a facet of USPKP. The following example illustrates this remark.

Example 3.1

Consider the following problem that has four jobs to be scheduled on a single machine with the data,

$$(r_{1,1},d_{1,1},\Delta_{1,1},w_{1,1})=(0,9,1,1),$$
$$(r_{2,1},d_{2,1},\Delta_{2,1},w_{2,1})=(0,9,1,1),$$
$$(r_{3,1},d_{3,1},\Delta_{3,1},w_{3,1})=(0,9,2,1),$$
$$(r_{4,1},d_{4,1},\Delta_{4,1},w_{4,1})=(0,9,1,1),$$

and has the duration [3,5] blocked-off on the single machine. The available time slots are numbered $k=1, \ldots, 7$ as shown:

Machine 1 Slots		1	2	3			4	5	6	7
Time	0	1	2	3	4	5	6	7	8	9

An optimal solution $\bar{x}$ to $\overline{\text{USP}}$ is given by

$$\bar{x}_{11}=1,\ \bar{x}_{22}=1/2,\ \bar{x}_{23}=1/2,\ \bar{x}_{35}=1,\ \bar{x}_{43}=1/2,\ \bar{x}_{47}=1/2,\ \text{and}\ \bar{x}_{ik}=0\ \text{otherwise.} \tag{3.6}$$

For $q=1$, we have that $T_q=\{1, 2, 3\}$ and the surrogate inequality of the type (Equation 3.4) is

$$x_{11z}+x_{12}+x_{13}+x_{21}+x_{23}+x_{32}+2x_{33}+2x_{42}+2x_{43}\leq 3. \tag{3.7}$$

Consider the corresponding continuous solution $\bar{y}$ translated to y-variables as given by $\bar{y}_{11}=1$, $\bar{y}_{12}=0$, $\bar{y}_{21}=1$, $\bar{y}_{22}=0$, $\bar{y}_{31}=0$, $\bar{y}_{41}=1/2$, and $\bar{y}_{42}=1/2$. Corresponding to Equation 3.7, the knapsack constraint of USPGAP is $y_{11}+y_{21}+2y_{31}+2y_{41}\leq 3$. From this knapsack constraint, we can obtain an extended minimal cover inequality, $y_{11}+y_{21}+y_{31}+y_{41}\leq 2$, which is a facet of the corresponding knapsack polytope and which deletes the solution $\bar{y}$. By Proposition 1, we have the following facet of USPGAP given by Equation 3.5, which is a valid inequality for Model USP: $x_{11}+x_{12}+x_{13}+x_{21}+x_{23}+x_{32}+x_{33}+x_{42}+x_{43}\leq 2$. Note that this inequality cuts off the fractional optimal solution (Equation 3.6) of $\overline{\text{USP}}$. Furthermore, augmenting USP by adding this cutting plane produces the integer optimal solution via $\overline{\text{USP}}$ given by $x_{11}=1$, $x_{24}=1$, $x_{36}=1$, $x_{43}=1$, and $x_{ik}=0$ otherwise.

3.3.3 A Class of Valid Inequalities from the Clique Inequality of the Set Packing Polytope

In this section, we consider a class of strong valid inequalities that are derived from the facets of the set packing polytope. For the convenience of presentation, let us define some notation. Let the constraints Equations 3.1 and 3.2 of USPP be numbered consecutively, and accordingly, let the index set M_G denote indices of the GUB constraints (Equation 3.1), and similarly, let M_P denote indices of the set packing constraints (Equation 3.2). Note that each constraint $i\in M_G$ corresponds to a job i and each constraint $k\in M_P$ corresponds to a slot k so that we can assume $M_G\equiv\{1, \ldots, n\}$ and $M_P\equiv\{n+1, \ldots, n+|K|\}$. Define $M_{\text{USP}}\equiv M_G\cup M_P$. Let us also refer to the variables as being singly subscripted and given by x_j for j belonging to an index set N_{USP}. Furthermore, denote by N_i the set of indices of variables contained in constraint $i\in M_{\text{USP}}$, and conversely, let M_j denote the index set of constraints in M_{USP} that include the variable x_j for each $j\in N_{\text{USP}}$. Note that with this revised compact notation, Model USP can be simply restated as

USP:

$$\text{Minimize}\left\{\sum_{j\in N_{\text{USP}}} c_j x_j : \sum_{j\in N_i} x_j = 1,\ \forall i\in M_G,\right.$$

$$\left.\sum_{j\in N_i} x_j \leq 1,\quad \forall i\in M_P,\ x_j\in\{0,1\},\ \forall j\in N_{\text{USP}}\right\}.$$

Similarly, we may restate Model USPP as follows:

USPP:

$$\text{Maximize}\left\{\sum_{j\in N_{\text{USP}}} \bar{c}_j x_j : \sum_{j\in N_i} x_j \leq 1,\ \forall i\in M_G,\right.$$

$$\left.\sum_{j\in N_i} x_j \leq 1,\quad \forall i\in M_P,\ x_j\in\{0,1\},\ \forall j\in N_{\text{USP}}\right\}.$$

For simplicity of notation, let us rewrite Model USPP as

USPP:

$$\text{Maximize}\left\{\bar{c}^T x : Ax\leq e,\ x_j\in\{0,1\},\ \forall j\in N_{\text{USP}}\right\},$$

where $e=\{1, \ldots, 1\}^T$. Define the *intersection graph* $G=(V, E)$ of the matrix A to have one node for each column of the matrix A, and one arc for every pair of nonorthogonal columns of A, i.e. $(i, j)\in E$ if and only if the corresponding columns a_i, a_j of A satisfy $a_i^T a_j\geq 1$. Let A_G be the node-arc incidence matrix of G, and denote by VP the weighted *node (vertex) packing* problem on the intersection graph G [13]. Define a *clique* as a maximal complete subgraph of G. Then, we have the following well-known class of facets for CHUSPP [18].

Proposition 3.2. The inequality $\sum_{j\in H} x_j \leq 1$ is a facet of the set packing polytope if and only if H is the node set of a clique of G.

The size of the associated intersection graph G is too large to permit the generation of all the clique facets of VP. However, the special structure of the matrix A lends itself to a convenient decomposition of its intersection graph G for identifying all the cliques of G. Consider a subgraph G^q of G, corresponding to the block interval $q\in Q$. Hence, $G^q=\{V^q, E^q\}$, where $V^q=\{\text{columns } j \text{ of } A: j\in N_{n+k} \text{ for some slot } k\in T_q\}$, and $E^q=\{(i, j)\in E: i, j\in V_q\}$. Note that $V=\cup_{q\in Q} V^q$. Let R_q be a clique of the subgraph G^q. Then,

$$\sum_{j\in R_q} x_j \leq 1 \tag{3.8}$$

is a valid inequality for USPP. Furthermore, each GUB constraint of Model USPP, restated below, corresponds to a complete subgraph of G:

$$\sum_{j\in N_i} x_j \leq 1, \quad \forall i \in M_G. \tag{3.9}$$

Proposition 3.3. Every clique facet of USPP belongs either to Equation 3.8 or to Equation 3.9.

Proof: Since the matrix A has a block diagonal structure corresponding to constraints in M_G as well as to those in M_P, for any $j_1\in V^{q_1}$, $j_2\in V^{q_2}$, where $q_1\neq q_2$, we have $a_{j_1}{}^T a_{j_2}\geq 1$ if and only if j_1 and j_2 appear in some common constraint in M_G, and $a_{j_1}{}^T a_{j_2}=0$, otherwise. Hence, the cliques of G, except for those of Equation 3.9, can only exist within some subgraph G^q, $q \in Q$. □

Proposition 3.3 asserts that we need to search only the subgraphs G^q, $q\in Q$, in order to generate all the clique facets of CHUSPP, other than those that already exist within Equation 3.9. Although we have reduced the search space for generating all the clique inequalities, finding all the cliques within each G^q still remains a hard problem. However, we can devise a simple separation procedure to generate a strong valid cutting plane inequality that deletes off a fractional optimal solution to USP as follows.

3.3.4 Separation Procedure

Let $\bar{x}$ be a fractional optimal solution to $\overline{\text{USP}}$, so that $F=\{j: \bar{x}_j \text{ is fractional}\}\neq\varnothing$. If any $\bar{x}_j=1$, all the vertices that are adjacent to the vertex j in the intersection graph cannot be packed, i.e. $\bar{x}_i=0$, $\forall(i, j)\in E$. Hence, we can further reduce the size of the subgraph G^q for finding a clique that generates a strong cutting plane inequality. In order to implement this reduction, we first try to find a complete subgraph on the vertex set corresponding to the set of fractional variables $\bar{x}_j$ in the block q. If we find such a complete subgraph with vertex set H such that $\sum_{j\in H\cap F}\bar{x}_j>1$, then we will have identified a clique inequality $\sum_{j\in H}\bar{x}_j\leq 1$ that cuts off $\bar{x}$. Subsequently, we can extend H to a maximal complete subgraph (clique R_q) of the subgraph G^q by examining the zero variables in V^q, and hence obtain a facetial cutting plane for USPP.

The following example illustrates the foregoing process for generating a strong valid cutting plane inequality from the facets of the set packing polytope.

Example 3.2

Consider the following data for a problem that has three jobs to be scheduled on two machines:

$$(r_{1,1},d_{1,1},\Delta_{1,1},w_{1,1}) = (0,9,2,1), \quad (r_{1,2},d_{1,2},\Delta_{1,2},w_{1,2}) = (0,9,2,1),$$
$$(r_{2,1},d_{2,1},\Delta_{2,1},w_{2,1}) = (0,9,3,1), \quad (r_{2,2},d_{2,2},\Delta_{2,2},w_{2,2}) = (0,9,3,1),$$
$$(r_{3,1},d_{3,1},\Delta_{3,1},w_{3,1}) = (0,9,4,1), \quad (r_{3,2},d_{3,2},\Delta_{3,2},w_{3,2}) = (0,9,4,1).$$

The machine blocked-off duration times are as follows: Machine 1: [3,5], Machine 2: [4,6]. The available time slots are numbered $k = 1, \ldots, 14$ as shown:

Machine 1 Slots	1	2	3			4	5	6	7
Machine 2 Slots	8	9	10	11			12	13	14

Time	0	1	2	3	4	5	6	7	8	9

An optimal solution $\bar{x}$ to $\overline{\text{USP}}$ is given by $\bar{x}_{1,9}=1/2$, $\bar{x}_{1,11}=1/2$, $\bar{x}_{2,3}=1/2$, $\bar{x}_{3,7}=1/2$, $\bar{x}_{3,11}=1/2$, and $\bar{x}_{i,k}=0$ otherwise. For this solution, variables $x_{1,9}$, $x_{1,11}$, $x_{3,11}$ are fractional in block $q=3$, where $T_q=\{8, 9, 10, 11\}$. However, the indices of these variables $\{x_{1,9}, x_{1,11}, x_{3,11}\}$ themselves yield a complete subgraph H of the subgraph of G^q that is induced by the fractional variables of V^q, for $q=3$. The node set of the extended complete subgraph of G^q, which includes H in its node set, corresponds to the indices of the variables $\{x_{1,9}, x_{1,10}, x_{2,10}, x_{2,11}, x_{3,11}\}$. Hence, the corresponding clique inequality of Equation 3.8 is $x_{1,9}+x_{1,10}+x_{2,10}+x_{2,11}+x_{3,11}\leq 1$, which cuts off the fractional optimal solution $\bar{x}$ to $\overline{\text{USP}}$. In addition, by augmenting USP with this additional clique inequality, we obtain an integer optimal solution for $\overline{\text{USP}}$, given by $x_{1,7}=1$, $x_{2,3}=1$, $x_{3,11}=1$, and $x_{ik}=0$ otherwise, which therefore solves Problem USP.

3.4 Solution Procedures for Problem USP

We begin below by devising a simple truncated implicit enumeration heuristic for Problem USP that appears to work quite well in practice. Following this, we will enhance this heuristic using both primal and dual approaches, in order to derive better quality solutions. The proposed truncated implicit enumeration heuristic, denoted HUSP, exploits the GUB structure of the problem when extending a partial solution, and it attempts to generate a good quality feasible solution in a constructive depth-first enumeration fashion, truncating the process once such a feasible solution has been found. We use the following notation:

PS: Partial solution comprised of an index set j such that $x_j=1$.

P_i: A set of variables corresponding to an unassigned target i, that are forced to zero due to the partial schedule *PS*; that is, $P_i=\{j: j \in N_i \cap N_s, \text{ for } s \in M_k \cap M_P, \text{ for } k \in PS\}$.

$PBS(i)$: An index set of variables x_j for $j \in (N_i - P_i)$, corresponding to an unassigned target $i \in M_G$ such that putting $x_j=1$ is known (by a previous attempt) to cause an infeasible partial schedule.

3.4.1 Heuristic HUSP

Intialization. Suppose that the variables x_j are arranged in descending order of $\bar{c}_j/|M_j|$, where $\bar{c}_j=(\theta - c_j)$, $\forall j \in N_{\text{USP}}$ and where $\theta \equiv 1+\sum_{j\in N} \max_{k\in S_i}\{c_{ik}\}$, and suppose that the targets are ordered and reindexed in descending order of the difference $\max_{k\in S_i}\{c_{ik}\}-\min_{k\in S_i}\{c_{ik}\}$ (see Remark 3.4). Level t of the depth-first enumeration tree that is constructed then corresponds to target t, for $t=1, \ldots, n$. Select target $t=1$, and set $PS=\varnothing$ and $PBS(i)=\varnothing$ for all $i \in M_G$.

Step 1. Determine the set P_t.

Step 2. Find the index j_t of target t such that

$$j_t = \arg\max\left\{\frac{\bar{c}_j}{|M_j|} : j \in [N_t - (P_t \cup PBS(t))]\right\}.$$

If j_t exists, then set $PS \leftarrow PS+\{j_t\}$, and go to Step 4. Otherwise, go to Step 3.

Step 3 (backtracking). If $t=1$ (or equivalently, if $PS=\varnothing$), then the given instance of Problem USP is infeasible. Otherwise, perform the following steps:

3.1. Set $PBS(t)=\varnothing$.
3.2. Replace $PS \leftarrow PS-\{j_{t-1}\}$.
3.3. Replace $PBS(t-1)\leftarrow PBS(t-1)+\{j_{t-1}\}$.
3.4. Replace $t\leftarrow t-1$.
3.5. Go to Step 2.

Step 4 (stopping criterion). If $t=|M_G|\equiv n$, then stop; we have found a feasible solution to Problem USP. Otherwise, set $t\leftarrow t+1$ and go to Step 1.

Remark 3.4. By its enumerative process, Heuristic HUSP will find a feasible solution in a finite number of steps, if problem USP is feasible. Note that in order to enhance its performance, we can rearrange the set of targets in the search process according to some rule that reflects the problem characteristics as governed by release times, due-dates, processing times, and the structure of the cost function. A set of possible alternative criteria for rearranging the targets is given in Lee [13]. Based on some preliminary experience, we adopted the rule given in the Initialization step of HUSP, following general heuristic guidelines used for set packing/partitioning. Note that the ordering of the variables is motivated by the fact that Problem USP can be equivalently stated as a set-packing problem, which is formulated in USPP above.

Despite the effectiveness of Heuristic HUSP, as will be demonstrated in the computational experience section, this procedure can be further enhanced to yield very close to optimal solutions by using the linear programming relaxation $\overline{\text{USP}}$ of Problem USP, which happens to provide a very tight lower bound. We chose to obtain quick lower and upper bounds on USP via USP by exploiting its special structure within the context of a Lagrangian relaxation approach (see Fisher [5]). To describe this approach, let u_i, $i\in M_G$ be the Lagrange multipliers associated with the GUB constraints of Model USP. Then, the corresponding Lagrangian relaxation of Model USP, denoted by LRUSP(u), can be constructed as follows:

$$\text{LRUSP}(u)\text{: Minimize} \sum_{i\in M_G}\sum_{j\in N_i}(c_j-u_i)x_j+\sum_{i\in M_G}u_i,$$

$$\text{subject to} \sum_{j\in N_i}x_j\le 1,\quad \text{for } i\in M_P, \tag{3.10}$$

$$x_j\in\{0,1\},\qquad \forall j\in N_{\text{USP}}. \tag{3.11}$$

For any $u\in R^{|M_G|}\equiv R^n$, we have $v(\text{LRUSP}(u))$ is a lower bound on $v(\text{USP})$, where $v(P)$ denotes the optimal value of any given problem P. The best choice of u is one that yields the greatest lower bound, and is given as an optimal solution to the following associated Lagrangian dual problem:

$$\text{LDUSP: } \underset{u\in R^n}{\text{Maximize}}\ v\big(\text{LRUSP}(u)\big).$$

Because the constraints set (Equation 3.10) has an interval matrix structure, as well as a block-diagonal structure, LRUSP(u) is a separable shortest path problem on an acyclic digraph G. Consequently, for any $u\in R^n$, every feasible extreme point solution to LRUSP(u) is integer-valued. This integrality condition implies that $v(\overline{\text{USP}})=v(\text{LDUSP})$.

In order to solve the Lagrangian dual problem LDUSP(u), we used a conjugate subgradient optimization algorithm described in Sherali and Ulular [23], called the average direction strategy (ADS), which conjugates the direction d^l at any iteration l as follows.

$$d^l=\xi^l \quad\text{and}\quad \left\{d^l=\xi^l+\frac{\|\xi^l\|}{\|d^{l-1}\|}d^{l-1}\right\},\text{ for } l\ge 2, \tag{3.12}$$

where ξ^l is a subgradient of the Lagrangian dual function at the lth iterate u^l, and $\|\bullet\|$ denotes the Euclidean norm. Note that that the direction d^l simply bisects the angle between ξ^l and d^{l-1}, and in this sense, is an "average direction". Sherali and Ulular [23] discuss the related theory that establishes the convergence of this scheme to an optimum for LDUSP and show that this direction has an empirical computational advantage over the procedures of Held, Wolfe, and Crower [12] and Camerini, Fratta, and Maffioli [2]. The choice of step sizes to be used in the context of subgradient optimization is a very crucial issue. In practice, the selection of a proper step size λ^l that enables a near-optimal incumbent solution to be obtained in a moderate number of iterations requires elaborate fine-tuning. Similar to step sizes prescribed by Held et al. [12], a suitable step size that can be used in concert with the ADS direction strategy is given by

$$\lambda^l = \beta^l \frac{(\bar{z} - z^l)}{\|d^l\|^2}, \tag{3.13}$$

where $\bar{z}$ is the current best known objective value for Problem USP, and β^l is a suitable parameter. Following the experience of Sherali and Ulular [23], we used the following block-halving strategy in order to accelerate the procedure. In this strategy, the iterations are partitioned into blocks based on an integer parameter N_2 such that at the top of each block when $l\,\mathrm{mod}(N_2) = 1$, we compute the step length λ^l using Equation 3.13 and with $\beta^l = \bar{\beta}/2^{\lfloor l/N_2 \rfloor}$, for a given $\bar{\beta} \in (0, 1)$.

Then, within each block, the step size is held constant, except that it is halved from its previous value, whenever the dual objective fails to improve over some N_3 consecutive iterations. Additionally, each time the step length needs to be revised (recomputed via Equation 3.13, or halved) over the N_1 iterations, the dual iterate is reset to the incumbent solution before using this modified step length. Typical parameter values prescribed are $N_1 = 200$, $N_2 = 75$, $N_3 = 5$, and $\bar{\beta} = 0.75$.

We will refer to the combinations of the ADS conjugate subgradient strategy and the foregoing block-halving step size strategy as ADSBH. For the purpose of comparison, we attempted various combinations of conjugate and pure subgradient direction-finding strategies along with various step-size rules. In our experience (see Lee [13] for details), we found ADSBH to dominate all the combinations tested. Among the latter alternatives, however, a competitive combination turned out to be one that employs the subgradient direction $d^l = \xi^l$ along with the block-halving step-size strategy.

3.5 A Branch-and-Bound Algorithm

In this section, we develop a branch-and-bound algorithm based on the bounds generated via a Lagrangian relaxation procedure, and a depth-first-search rule for selecting an active node in the branching process. Below, we consider various algorithmic strategies related to the Lagrangian dual procedure, preprocessing rules, logical tests based on reduced costs, the development of an enumeration tree, the node selection rule employed, and the fathoming process.

Notationally, let x^b denote the current best primal feasible solution found thus far, and let UB denote the upper bound corresponding to x^b. In addition, at each node r of the branching tree, let u^r denote the best dual feasible solution found to the continuous relaxation of the node subproblem, and let LB denote the lower bound corresponding to u^r.

3.5.1 Preprocessing Rules

Preprocessing rules can be very useful in reducing the number of variables and constraints. These rules employ simple row and column scans to identify redundant variables that can be fixed at 0 or 1, and to modify matrix coefficients. Below, we describe some simple preprocessing rules for detecting infeasibilities or reducing the size of the problem by identifying variables that can be fixed at 0 or 1. These rules are applied to the initial problem as well as to each node of the branching tree, and can be very effective in solving the problem.

Problem decomposition rule: Suppose that all the jobs are sorted in ascending order *of* $\min_{j\in M}\{r_{i,j}\}$. Define $\bar{h} = \min\{h: \max_{j\in M}\{d_{i,j}\} \leq \min_{j\in M}\{r_{h+1,j}\}$, for $1 \leq i \leq h\}$, where $r_{n+1,j} \equiv \infty$. Then, we can decompose Problem USP into separable problems over the set of jobs N_1 and N_2, where $N_1 = \{1, \ldots, \bar{h}\}$ and $N_2 = N - N_1$. Treating N_2 as the original set of jobs N, this operation can then be repeated with N_2. In this manner, we can obtain a set of separable problems that can be solved independently to obtain an optimal solution.

Problem reduction rule (variable fixing): If $|N_i| = 1$ for any $i \in M_G$, then we can set the corresponding variable $x_j = 1$, delete x_j in the variable set, delete the set of constraints from the model, and also put $x_l = 0$ for all $l \in N_t$, $t \in M_j$, $l \neq j$. In the same spirit, if for some $j \in N_{\text{USP}}$, there exists $i \in M_G$, $j \notin N_t$, such that $\bigcup_{l\in Ni} (M_\lambda \cap M_P) \subseteq M_j \cap M_P$, then we can set $x_j = 0$ and delete this variable from the model. In other words, if putting $x_j = 1$ creates a conflict for all possible slot assignments for some job i, then we can set $x_j = 0$.

An implication of the reduced costs: At each node, we implement a variable-fixing procedure prior to any further branching. That is, at the root node, for example, we can delete from the model any variable $j \in N_{\text{USP}}$ for which the reduced cost exceeds $UB - LB_0$, since $x_j = 1$ cannot be optimal. Similarly, we can apply such a variable-fixing procedure at each node based on the reduced costs corresponding to the dual variables u^r obtained for the associated node subproblem, in order to delete any variable from the corresponding subproblem whose reduced cost exceeds the bound gap ($UB - LB_r$).

3.5.2 Development of the Enumeration Tree

We implement a depth-first-search rule to construct the enumeration tree. Accordingly, we denote by node r, the currently active node at the rth level of the enumeration tree. At any such level $r > 0$ of the enumeration tree, the basic branching step selects a GUB constraint $i(r)$ corresponding to some job $i(r)$ on which we have not as yet branched at a predecessor node on the chain to the root node, and creates a node for each $j \in N_{i(r)}$. We consider two rules for selecting a constraint. First, we implement *a fixed-order* rule for sequencing the choice of the branching constraints. That is, at the root node, a branching sequence is determined *a priori* in descending order of the values $|u_i^0|$, $i \in M_G$, motivated by the marginal cost interpretation of dual values. As a second rule, we *dynamically* select a branching constraint $i(r)$ at the rth level of the enumeration tree such that $i(r) \in \arg\max\{|u_i^r|$: $i \in$ (unbranched constraints in the chain from node r to the root node)} and where u^r is the best dual solution obtained for the most recent Lagrangian dual problem solved at level r. In Section 3.6, we report on computational results obtained using these rules for selecting branching constraints in the context of a branch-and-bound algorithm. Having selected a branching constraint $i(r)$ at the rth level of the branching tree, we create a node for each $j \in N_{i(r)}$, fixing $x_j = 1$ and $x_t = 0$ for all $t \in (N_{i(r)} - j)$ at each such node. These nodes, thus generated for level $(r+1)$, are stored in a linked list of open nodes, denoted by Λ, in descending order of the reduced costs corresponding to the best dual solution obtained for the Lagrangian dual problem solved at the immediate predecessor node at level r, and are explored further in a Last-In-First-Out (LIFO) fashion.

3.5.3 Subproblem Generation

Each node of the branch-and-bound tree defines a new scheduling problem of reduced size. Consider a partial solution $FS(r)$ generated at node r at the rth level of the enumeration tree, obtained by fixing some variable $x_j = 1$, $j \in N_{i(s)}$, for each level $s < r$ of the enumeration tree.

Specifically, let $FS(r)$ be the set of indices of variables x_j, $j \in N_{i(r)}$, $s = 0, \ldots, r-1$, that are restricted to unity on the unique chain of the enumeration tree traced from node r to the root node. At the rth level of the enumeration tree, we generate for level $(r+1)$, a set of problems P_s^{r+1}, obtained by fixing $x_t = 1$, $\forall t \in FS(r)$ for each $s \in N_{i(r)}$.

Note that each problem P_s^r, for each $s \in N_{i(r-1)}$, has the same structure as that of Problem USP, but is of a reduced size. Problem P_s^r can be further reduced by explicitly deleting zero variables and resulting trivial constraints. However, this reduction can be efficiently handled by implicitly deleting zero arcs and isolated nodes of the underlying acyclic digraph associated with Problem USP (see Remark 3.1).

Obviously, some problem P_s^r is infeasible if $(M_P \cap M_s) \cap (\bigcup_{j \in FS(r)} M_j) \neq \emptyset$. These trivial infeasible problems are not stored in the list Λ of open nodes. Instead of storing an explicit representation of the formulation P_s^r, or for that matter, instead of even storing an explicit representation of Λ, we record minimal information that can be used to recover the list Λ, as well as recover for each node in the list Λ, the level of the node, the index set of the fixed variables at the node, and the (near) optimal dual solution of the immediate predecessor node. Hence, when we (conceptually) retrieve an open node from Λ, we can efficiently construct an explicit representation of the corresponding subproblem. Moreover, note that the foregoing preprocessing procedures and the variable-fixing procedure are applied to each new reduced node subproblem created.

3.5.6 Node Selection Rule and Fathoming Process

We conduct a standard LIFO/depth-first branch-and-bound process on the list Λ of open nodes. Note that in this process, for each explored unfathomed node, of which there is one at each level, the corresponding (near) optimal dual solution (u_i^r, $i \in M_G$ for each r) is saved so that it can be used as an advanced starting solution for the immediate descendant nodes.

3.5.7 Algorithm USPBB

We now present the proposed detailed algorithmic steps for finding exact and/or approximate solutions for Problem USP using the foregoing branch-and-bound procedure. This algorithm requires a specification of the following parameters:

1. ε = guaranteed optimality parameter. Hence, whenever $(UB - LB_r)/UB \leq \varepsilon$, we fathom node r. If $\varepsilon = 0$, the algorithm is an exact procedure for finding an optimal integer solution to Problem USP. On the other hand, if $0 < \varepsilon < 1$, the algorithm tries to find a feasible solution such that the optimal objective value of Problem USP lies in the interval $[(1-\varepsilon)UB, UB]$.
2. τ_{BB} = the maximum allowed run-time. If the total run-time of the branch-and-bound algorithm, checked at the top of the loop, exceeds τ_{BB}, then the algorithm is terminated, and the current best feasible solution is taken as the prescribed solution.
3. τ_{HUSP} = the maximum allowed run-time for the heuristic HUSP, at the end of which, this procedure is aborted.

Initialization: Initialize the enumeration tree level of the root node as $r = 0$. Determine an initial feasible solution x^b for Problem USP via the heuristic HUSP of Section 3.4, and let UB be its objective value. Also select an initial Lagrangian dual vector u^0 as $u_i^0 = \min_{j \in N_i}(c_j)$, $\forall i \in M_G$, and denote by $LB_0 = \sum_{i \in M_G} u_i^0$ the incumbent solution value of LDUSP. If $(UB - LB_r)/UB \leq \varepsilon$, for some tolerance $0 < \varepsilon < 1$, terminate with x^b as the prescribed solution. Else, proceed to Step 1.

Step 1 (root node). Solve the Lagrangian dual problem (LDUSP) of Model USP via the conjugate subgradient optimization method to find a revised (near) optimal dual solution u^0. Also, denote by LB_0 the (near) optimal objective function value of LDUSP. If $(UB - LB_0)/UB \leq \varepsilon$, then stop with x^b as an ε-optimal solution. Otherwise, execute the variable-fixing procedure using the reduced cost vector. Put $FS(r) = \emptyset$, and proceed to Step 2.

Step 2 (generation of subproblems). Select the branching GUB constraint or job $i(r)$ as in Section 3.5.2. Let $J_{i(r)} = \{j \in N_{i(r)}: M_j \cap M_l = \emptyset, \forall l \in FS(r)\}$. (Note that if $J_{i(r)} = \emptyset$, then the subproblem at the preceding level would have been infeasible, even in the continuous sense.) Construct subproblems P_s^{r+1} for all $s \in J_{i(r)}$, and add these subproblems to the linked list Λ of open nodes, according to a descending order of the reduced costs on stated above. Increment r by one and proceed to Step 3.

Step 3 (run-time check). Check the accumulated run-time. If this time exceeds the limit τ_{BB}, then stop. Otherwise, proceed to Step 4.

Step 4 (problem selection and relaxation). Select the last subproblem P_t^r stored in the list Λ, and delete it from Λ. Let $FS(r)=FS(r-1)+\{t\}$. Solve the Lagrangian dual of problem PS_t^r using an advanced starting dual solution available from its immediate predecessor node in the branch-and-bound tree. Let u^r be the best dual solution of objective value LB_r. If $(UB-LB_r)/UB\leq\varepsilon$, then fathom this node and go to Step 6. Otherwise, proceed to Step 5.

Step 5 (new incumbent solution by heuristic HUSP). Let $\bar{x}^r$ denote an optimal solution to the Lagrangian relaxation subproblem of Problem P_t^r, corresponding to the (near) optimal dual solution u^r obtained at Step 4. Denote $M_G^+=\{i\in M_G: \sum_{j\in N}\bar{x}_j^r=1\}$ and let $M_G^-=M_G-M_G^+$. If $M_G^-=\emptyset$, then $\bar{x}^r$ is an optimal solution for Problem $\bar{P}_t^r$, the linear programming relaxation of P_t^r. In addition, if $\bar{x}^r$ is integral, it is optimal to P_t^r, and so, if $LB_r<UB$, then let $UB=LB_r$ and $x_b=\bar{x}^r$, and go Step 6. Otherwise, if $\bar{x}^r$ is fractional, order the variables separately within each of the two sets M_G^+ and M_G^- in ascending order of $c_j/|M_j|$ and concatenate these in the stated order. Then, implement Heuristic HUSP using this ordering of the jobs, and starting with an advanced starting partial solution $PS=\{j: \bar{x}_j^r=1 \text{ for } j\in N_i,\ i\in M_G^+\}$, where the jobs in PS are arranged as they are in M_G^+. If HUSP terminates before the time limit τ_{HUSP} with an indication that the problem is infeasible, then fathom this node and go to Step 6. If the objective value of the solution produced by HUSP is less than UB, then replace x^b by this improved heuristic solution to USP, and set $UB=cx^b$. If $(UB-LB_r)/UB\leq\varepsilon$, then fathom the current node and go to Step 6. Otherwise, return to Step 2.

Step 6 (Fathoming step and termination check). Replace $FS(r)\leftarrow FS(r)-\{t\}$. If $\Lambda=\emptyset$, then stop; we have found an ε-optimal solution x^b to problem USP. Return to Step 3.

Remark 3.5. Note that in practice, if no feasible solution has been found in a given time limit, then the available partial solution generated by the proposed algorithm is used to prescribe imminent target scheduling decisions, and the horizon is then advanced and the problem is resolved.

3.6 Computational Results

We now report on our computational experience in implementing Algorithm USPBB for solving large-scale instances of Problem Naval Battle Group Problem (NBG) with size characteristics given in Table 3.1. The data for these problems are generated using guidelines provided by the Naval Surface Warfare Center (NSWC), Dahlgren, Virginia. Accordingly, the release times $r_{i,j}$ were generated uniformly on the interval [0, 50], the processing times $\Delta_{i,j}$ were generated uniformly over [1,6], and the due-dates $d_{i,j}$ were computed as $d_{i,j}=r_{i,j}+\Delta_{i,j}+\delta_{i,j}$, where $\delta_{i,j}$ are generated uniformly over [1,16]. The number of blocked-off durations, of unit lengths, were generated uniformly over [0, 10] for each machine, and were then distributed uniformly over the time horizon $[0, \max_{i,j}\{d_{i,j}\}]$. In addition, the weights $w\equiv(w_{ik}, k\in S_i, i\in N)$ were taken as either w_1 or w_2 specified below, again using NSWC guidelines based on some internal unpublished studies).

$$w_{ik}^{1}=\left[\frac{10^4}{(d_{i,\mu_k}-r_{i,\mu_k})^2}\right],\quad w_{ik}^{2}=\left[\frac{\max_{p,q}(d_{p,q}-r_{p,q})^2}{(d_{i,\mu_k}-r_{i,\mu_k})^2}\right],\forall k\in S_i,\ i\in N. \tag{3.12}$$

Recall that the objective function of Problem USP seeks to minimize $\sum_{i\in N}\sum_{k\in Si}w_{ik}\tau_k x_{ik}$. Hence, using Equation 3.12, preference is given to assigning target i to that slot k on illuminator μ_k, for which not

TABLE 3.1 Size Characteristics of the Test Problems

	Number of Targets	Number of Illuminators	Number of Variations of Model USP	Number of Rows of Model USP
Problem 1	30	6	1007	157
Problem 2	40	8	1807	241
Problem 3	40	5	1561	351
Problem 4	45	5	1868	369
Problem 5	45	7	2356	454
Problem 6	50	5	1152	177
Problem 7	50	5	2148	387
Problem 8	50	5	2095	376
Problem 9	55	7	3077	504
Problem 10	60	8	3679	563

only does that engagement occur early in the time horizon, but also the corresponding availability time-window duration $(d_{i,\mu_k} - r_{i,\mu_k})$ is relatively large. The latter aspect is important in that if the strike is planned early in a relatively large time-window interval, then a possible miss or an unscheduled reallocation decision provides sufficient opportunity for using the same machine on a subsequent attempt. Both the choices of weights in Equation 3.12 reflect this bias, using either an absolute or a relative time-window interval term.

The computational times reported are all batch processing times in CPU seconds on an IBM 3090 Series 300E computer, with coding in PASCAL, and include model generation and all input and output times. For computational effectiveness, based on some preliminary computations, the maximum number of subgradient iterations for the Lagrangian dual problem at the root node was set at 300, the maximum number of iterations at the other enumeration nodes was set at 50, and the maximum run-times tolerances were set at $\tau_{BB}=1$ sec and $\tau_{\text{HUSP}}=1$ sec.

We now explore the effect of the various implemented strategies on the performance of Algorithm USPBB. Tables 3.2 and 3.3 shows computational results for Algorithm USPBB with the *fixed-order branching rule* and the *dynamic rule,* respectively, using the test problems of Table 3.1, along with the weight function w^1 as defined in Equation 3.12, and while using two optimality termination criteria, namely, $\varepsilon=0.01$ and $\varepsilon=0.005$. Similarly, Tables 3.4 and 3.5 shows computational results for Algorithm USPBB with the fixed-order branching rule and the dynamic branching rule respectively, using the test problems of Table 3.1, along with the weight function w^2 as defined in Equation 3.12, and while using two optimality termination criteria, namely, $\varepsilon=0.01$ and $\varepsilon=0.005$.

Observe that the conjugate subgradient method along with the block-halving step-length strategy quickly provides (near) optimal dual solutions to $\overline{\text{USP}}$, and serves to efficiently generate tight lower bounds $v(\overline{\text{USP}})$. In addition, the heuristic procedure described in Step 5 of Algorithm USPBB generates good quality upper bounds within a reasonable amount of computational effort. Consequently, these tight lower and upper bounds along with the preprocessing rules of Section 3.5.1, accelerate the performance of Algorithm USPBB. Moreover, since for most of the test problems, the gap between the initial lower and upper bounds is relatively small (3–5%), the branching strategies do not significantly affect the overall performance of Algorithm USPBB as evident from computational results. However, we would recommend the dynamic branching rule, which performs slightly better than the fixed-order rule. In addition, observe that Algorithm USPBB with the ADS strategy works slightly better than the algorithm with the HWC strategy for these test problems. For the most part, Algorithm USPBB finds a solution within the prescribed optimality tolerance of 1% in 3–10 sec of CPU time, which conforms with the desired performance limits set by NSWC.

TABLE 3.2 Computational Results for Algorithm USPBB using the Fixed-order Branching Rule Along with Weight Function *w*

	Best UB Found		*ν*(LDUSP) at Root Node		(Run-Time (seconds), No. of Nodes)		Optimality (ε=0.01)	
Problem	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH
Problem 1	62598	63123	62582.6	62594.6	(10.7, 82)	(7.8, 21)	99.9	99.2
Problem 2	92922	93372	92738.7	92577.6	(6.3, 0)	(3.4, 0)	99.8	99.1
Problem 3	45288	45330	45172.2	45175.1	(1.2, 0)	(1.5, 0)	99.7	99.7
Problem 4	48928	48500	48454.5	48456.3	(2.0, 0)	(3.3, 0)	99.0	99.9
Problem 5	50622	50253	50187.4	50194.3	(2.1, 0)	(2.2, 0)	99.1	99.9
Problem 6	203528	203928	201499.4	201080.9	(13.2, 0)	(30.0, 195)	99.0	99.0
Problem 7	57113	57130	56664.0	56671.9	(2.3, 0)	(2.7, 0)	99.2	99.2
Problem 8	58320	58699	57785.5	57795.1	(2.3, 0)	(30.0, 151)	99.1	98.5
Problem 9	52964	52864	52481.7	52469.8	(6.2, 0)	(6.5, 0)	99.1	99.3
Problem 10	63053	63271	63048.3	63033.9	(4.2, 0)	(4.6, 0)	99.9	99.6
	Best UB Found		***ν*(LDUSP) at Root Node**		**(Run-Time (seconds), No. of Nodes)**		**Optimality (ε=0.005)**	
Problem	**ADSBH**	**HWCBH**	**ADSBH**	**HWCBH**	**ADSBH**	**HWCBH**	**ADSBH**	**HWCBH**
Problem 1	62802	62838	62582.6	62594.6	(13.9, 140)	(14.2, 156)	99.7	99.6
Problem 2	92922	993372	92738.7	92577.6	(6.3, 0)	(3.4, 0)	99.8	99.1
Problem 3	45288	45330	45172.2	45175.1	(1.2, 0)	(1.5, 0)	99.7	99.7
Problem 4	48660	48500	48456.2	48456.3	(2.4, 0)	(3.3, 0)	99.6	99.9
Problem 5	50267	50253	50189.6	50194.3	(2.1, 0)	(2.2, 0)	99.8	99.9
Problem 6	203128	203928	201499.4	201080.9	(30.0, 190)	(30.0, 195)	99.0	98.6
Problem 7	56995	57000	56678.2	56680.9	(30.0, 59)	(14.2, 120)	99.4	99.5
Problem 8	58320	58386	57785.5	57795.1	(4.5, 1)	(130, 133)	99.5	99.0
Problem 9	51731	51799	52481.7	52478.8	(29.8, 150)	(18.9, 150)	99.5	99.5
Problem 10	63053	63271	63048.3	63033.9	(4.2, 0)	(4.6, 0)	99.9	99.6

Note: Optimality=100×[1−(Best UB−ν(LDUSP))/Best UB]

TABLE 3.3 Computational Results for Algorithm USPBB using the Dynamic Branching Rule Along with Weight Function *w*

	Best UB Found		*ν*(LDUSP) at Root Node		(Run-Time (seconds), No. of Nodes)		Optimality (ε=0.01)	
Problem	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH
Problem 1	62598	63123	62582.6	62594.6	(10.7, 82)	(7.8, 21)	99.9	99.2
Problem 2	92922	93372	92738.7	92577.6	(6.3, 0)	(3.4, 0)	99.8	99.1
Problem 3	45288	45330	45172.2	45175.1	(1.2, 0)	(1.5, 0)	99.7	99.7
Problem 4	48928	48500	48454.5	48456.3	(2.0, 0)	(3.3, 0)	99.0	99.9
Problem 5	50622	50253	50187.4	50194.3	(2.1, 0)	(2.2, 0)	99.1	99.9
Problem 6	203528	203128	201499.4	201080.9	(13.2, 0)	(30.0, 218)	99.0	99.0
Problem 7	57113	57130	56664.0	56671.9	(2.3, 0)	(2.7, 0)	99.2	99.2
Problem 8	58320	58699	57785.5	57795.1	(2.3, 0)	(30.0, 151)	99.1	98.5
Problem 9	52964	52864	52481.7	52469.8	(6.2, 0)	(6.5, 0)	99.1	99.3
Problem 10	63053	63271	63048.3	63033.9	(4.2, 0)	(4.6, 0)	99.9	99.6

(*Continued*)

TABLE 3.3 Computational Results for Algorithm USPBB using the Dynamic Branching Rule Along with Weight Function *w* (Continued)

	Best UB Found		*v*(LDUSP) at Root Node		(Run-Time (seconds), No. of Nodes)		Optimality (ε=0.005)	
Problem	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH
Problem 1	62802	62838	62582.6	62594.6	(13.9, 140)	(14.2, 117)	99.7	99.7
Problem 2	92922	92757	92738.7	92754.5	(6.3, 0)	(8.9, 210)	99.8	99.9
Problem 3	45288	45330	45172.2	45175.1	(1.2, 0)	(1.5, 0)	99.7	99.7
Problem 4	48660	48500	48456.2	48456.3	(2.4, 0)	(3.3, 0)	99.6	99.9
Problem 5	50267	50253	50189.6	50194.3	(2.1, 0)	(2.2, 0)	99.8	99.9
Problem 6	203128	203928	201499.4	201080.9	(30.0, 190)	(30.0, 195)	99.0	98.6
Problem 7	56995	57000	56678.2	56680.9	(30.0, 59)	(14.2, 120)	99.4	99.5
Problem 8	58320	58386	57785.5	57795.1	(4.5, 1)	(30.0, 150)	99.5	99.0
Problem 9	51731	51799	52481.7	52469.8	(27.5, 137)	(18.9, 150)	99.5	99.5
Problem 10	63053	63271	63048.3	63033.9	(4.2, 0)	(4.6, 0)	99.9	99.6

TABLE 3.4 Computational Results for Algorithm USPBB using the Fixed-order Branching Rule Along with Weight Function *w*

	Best UB Found		*v*(LDUSP) at Root Node		(Run-Time (seconds), No. of Nodes)		Optimality (ε=0.01)	
Problem	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH
Problem 1	584	584	578.7	580.9	(1.3, 0)	(1.1, 0)	99.5	99.1
Problem 2	869	859	855.5	854.9	(30, 212)	(3.5, 0)	98.5	99.5
Problem 3	1982	1983`	1972.4	1972.4	(1.4, 0)	(1.5, 0)	99.5	99.5
Problem 4	2089	2100	2072.4	2072.4	(14.4, 76)	(30.0, 97)	99.2	98.7
Problem 5	2098	2103	2094.3	2094.3	(2.2, 0)	(2.1, 0)	99.9	99.6
Problem 6	1008	1012	999.6	996.2	(14.1, 102)	(30.0, 112)	99.2	98.5
Problem 7	2567	2555	2541.2	2540.2	(24.7, 213)	(2.1, 0)	99.0	99.5
Problem 8	2539	2539	2515.3	2515.1	(3.0, 0)	(4.2, 0)	99.1	99.1
Problem 9	2171	2171	2147.5	2147.6	(18.6, 53)	(10.2, 1)	99.1	99.1
Problem 10	2640	2632	2623.8	2623.2	(4.6, 0)	(4.2, 0)	99.4	99.7

	Best UB Found		*v*(LDUSP) at Root Node		(Run-Time (seconds), No. of Nodes)		Optimality (ε=0.005)	
Problem	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH
Problem 1	584	584	580.9	580.9	(1.7, 13)	(1.1, 0)	99.5	99.5
Problem 2	864	859	855.5	854.9	(30, 323)	(3.5, 0)	99.1	99.5
Problem 3	1982	1983	1972.4	1972.4	(1.4, 0)	(1.5, 0)	99.5	99.5
Problem 4	2092	2101	2072.4	2072.4	(30.0, 163)	(30.0, 160)	99.1	98.7
Problem 5	2098	2103	2094.3	2094.3	(2.2, 0)	(2.1, 0)	99.9	99.6
Problem 6	1010	1012	999.6	996.2	(30.0, 209)	(30.0, 160)	99.2	98.5
Problem 7	2551	2555	2541.2	2540.2	(24.7, 198)	(11.6, 119)	99.6	99.5
Problem 8	2539	2539	2515.3	2515.1	(7.1, 1)	(7.5, 1)	99.5	99.1
Problem 9	2158	2167	2147.5	2147.6	(28.2, 66)	(30.0, 136)	99.5	99.1
Problem 10	2632	2632	2623.8	2623.2	(5.0, 0)	(4.2, 0)	99.7	99.7

TABLE 3.5 Computational Results for Algorithm USPBB using the Dynamic Branching Rule Along with Weight Function *w*

	Best UB Found		v(LDUSP) at Root Node		(Run-Time (seconds), No. of Nodes)		Optimality (ε=0.01)	
Problem	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH
Problem 1	584	584	578.7	580.9	(1.3, 0)	(1.1, 0)	99.5	99.1
Problem 2	869	859	855.5	854.9	(30, 212)	(3.5, 0)	98.5	99.5
Problem 3	1982	1983	1972.4	1972.4	(1.4, 0)	(1.5, 0)	99.5	99.5
Problem 4	2089	2100	2072.4	2072.4	(14.4, 76)	(30.0, 97)	99.2	98.7
Problem 5	2098	2103	2094.3	2094.3	(2.2, 0)	(2.1, 0)	99.9	99.6
Problem 6	1008	1012	999.6	996.2	(14.1, 101)	(30.0, 101)	99.2	98.5
Problem 7	2567	2555	2541.2	2540.2	(24.7, 208)	(2.1, 0)	99.0	99.5
Problem 8	2539	2539	2515.3	2515.1	(3.0, 0)	(4.2, 0)	99.1	99.1
Problem 9	2171	2171	2147.5	2147.6	(18.6, 53)	(10.2, 1)	99.1	99.1
Problem 10	2640	2632	2623.8	2623.2	(4.6, 0)	(4.2, 0)	99.4	99.7

	Best UB Found		v(LDUSP) at Root Node		(Run-Time (seconds), No. of Nodes)		Optimality (ε=0.005)	
Problem	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH	ADSBH	HWCBH
Problem 1	584	584	580.9	580.9	(1.6, 13)	(1.1, 0)	99.5	99.5
Problem 2	864	859	855.5	854.9	(30, 343)	(3.5, 0)	99.1	99.5
Problem 3	1982	1983	1972.4	1972.4	(1.4, 0)	(1.5, 0)	99.5	99.5
Problem 4	2092	2101	2072.4	2072.4	(30.0, 156)	(30.0, 74)	99.1	98.7
Problem 5	2098	2103	2094.3	2094.3	(2.2, 0)	(2.1, 0)	99.9	99.6
Problem 6	1010	1012	999.6	996.2	(30.0, 226)	(30.0, 191)	99.2	98.5
Problem 7	2551	2555	2541.2	2540.2	(24.7, 198)	(11.6, 119)	99.6	99.5
Problem 8	2539	2539	2515.3	2515.1	(7.1, 1)	(7.5, 1)	99.5	99.1
Problem 9	2158	2167	2147.5	2147.6	(28.2, 66)	(30.0, 136)	99.5	99.1
Problem 10	2632	2632	2623.8	2623.2	(5.0, 0)	(4.2, 0)	99.7	99.7

3.7 Summary and Discussion

We have constructed a strong zero-one integer programming formulation USP for an unrelated machine scheduling problem of minimizing the total weighted flow time, subject to specified job and machine availability constraints, and have devised an effective branch-and-bound algorithm that finds high quality solutions with an effort that is well within acceptable standards. This branch-and-bound algorithm exploits the inherent structures of the problem to efficiently generate tight lower and upper bounds via a Lagrangian relaxation approach. Using test problems that arise within the context of a naval battle-group engagement problem, we have demonstrated that this algorithm generates solutions guaranteed to be within 99% of optimality in 3–10 s of CPU time. The Naval Surface Warfare Center, Dahlgren, Virginia, has implemented this developed software on the Aegis type battleships.

Although for the problem instance of the cited application we did not find it necessary to generate additional valid inequalities, we have shown via examples how strong valid inequalities can be generated by exploiting the structure of Problem USP in order to tighten its linear programming relaxation. Such inequalities can be employed within the framework of the Lagrangian relaxation procedure to derive improved heuristic and exact procedures aimed at hard instances of this problem. In such an implementation, it would be advisable to dualize these additional inequalities along with the GUB constraints, in order to preserve the interval matrix structure of the subproblem constraints.

References

1. Akker, J.M., Hurkens, C.A.J, and Savelsbergh, M.W.P. 2000. Time-indexed formulations for machine scheduling problems: column generation, *INFORMS Journal on Computing*, 12(2), 111–124.
2. Camerini, P.M., Fratta, L., and Maffioli, F. 1975. On improving relaxation methods by modified gradient techniques, *Mathematical Programming Study*, 3, 26–34.
3. Cavalcante, G.C.B. Souza, C.C., Savelsberg, M.W.P., Wang, Y., and Wolsey, L.A. 2001. Scheduling projects with labor constraints, *Discrete Applied Mathematics*, 112, 27–52.
4. Eishafei, M., Sherali, H.D., and Smith, J.C. 2003. Radar pulse interleaving for multi-target tracking, *Naval Research Logistics*, 51, 72–94.
5. Fisher, M.L. 1981. The Lagrangian relaxation method for solving integer programming problems, *Management Science*, 27, 1–18.
6. Garey, M.R., and Johnson, D.S. 1977. Two-processor scheduling with start times and deadlines, *SIAM Journal of Computing*, 6, 416–426.
7. Garey, M.R., and Johnson, D.S. 1979. *Computer and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, New York.
8. Gottlieb, E.S., and Rao, M.R. 1990. The generalized assignment problem: valid inequalities and facets, *Mathematical Programming*, 46, 31–52.
9. Gottlieb, E.S., and Rao, M.R. 1990. (1, k)-Configuration facets for the generalized assignment problem, *Mathematical Programming*, 46, 53–60.
10. Grabowski, J., Nowicki, E., and Zdrzalka, S. 1986. A block approach for single machine scheduling with release dates and due dates, *European Journal of Operational Research*, 26, 278–258.
11. Hammer, P.L., Johnson, E.L., and Peled, U.N. 1975. Facets of regular 0-1 polytopes, *Mathematical Programming*, 8, 179–206.
12. Held, M., Wolfe, P., and Crowder, H.P. 1974. Validation of subgradient pptimization, *Mathematical Programming*, 6, 62–88.
13. Lee, Y. 1992. The Polyhedral Structure of Some Combinatorial Optimization Problems with Application to a Naval Defense Problem, Dissertation, Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University.
14. Lee, Y., and Sherali, H.D. 1994. Unrelated machine scheduling with time-window and machine-downtime constraints: an application to a naval battle-group problem, *Annals of Operations Research* (special issue on Applications in Combinatorial Optimization) 50, 339–365.
15. Leon, V.J., and Wu, S.D. 1992. On scheduling with ready-times, due-dates, and vacations, *Naval Research Logistics*, 39, 53–65.
16. Maskery, M., and Krishnamurthy, V. 2007. Network-enabled missile deflection: games and correlation equilibrium, *IEEE Transaction on Aerospace and Electronic System*, 43 (3), 843–853.
17. Nemhauser, H.L., and Wolsey, L.A. 1998. *Integer and Combinatorial Optimization*, Wiley, New York.
18. Padberg, M. 1973. On the facial structure of set packing polyhedra, *Mathematical Programming*, 5, 199–215.
19. Padberg, M., and Rinaldi, G. 1991. A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems, *SIAM Review,* 33, 60–100.
20. Potts, C.N. 1980. Analysis of a heuristic for one machine sequencing with release dates and delivery times, *Operations Research*, 28, 1436–1441.
21. Sarin, S.C., Ahn, S., and Bishop, A.B. 1988. An improved branching scheme for the branch and bound procedure of scheduling *n* jobs on *m* parallel machines to minimize total weighted flowtime, *International Journal of Production Research*, 26, 1183–1191.
22. Sherali, H.D., Lee, Y., and Boyer, D.D. 1995. Scheduling target illuminators in naval battle-group anti-air warfare, *Naval Research Logistics*, 42(5), 737–755.

23. Sherali, H.D., and Ulular, O. 1989. A primal-dual conjugate subgradient algorithm for specially structured linear and convex programming problems, *Applied Mathematics and Optimization*, 20, 193–223.
24. Simons, B., On Scheduling with Release Times and Deadlines, in Dempster M.A.H., et al. (eds.). 1982. *Proceedings of the NATO Advanced Study and Research Institute on Theoretical Approach to Scheduling Problems*, D. Reidel Co., Hingham, 75–88.
25. Simons, B. 1983. Multiprocessor scheduling of unit-time jobs with arbitrary release times and deadlines, *SIAM Journal of Computing*, 12, 294–229.
26. Simons, B., and Warmuth, M.K. 1989. A fast algorithm for multiprocessor scheduling of unit-length jobs, *SIAM Journal of Computing*, 18, 690–710.
27. Sousa, J.P., and Wolsey, L.A. 1994. A time index formulation on nonpreemptive single machine scheduling problem, *Mathematical Programming*, 54, 353–367.
28. Wolsey, L.A. 1997. MIP modeling of changeovers in production planning and scheduling problems, *European Journal of Operational Research*, 99, 154–165.
29. Zemel, E. 1989. Easily computable facets of the knapsack polytope, *Mathematical Programming*, 14, 760–764.

4

Multiple Criteria Optimization Models for Supplier Selection

A. Ravi Ravindran
Pennsylvania State University

Vijay Wadhwa
Pennsylvania State University

4.1 Supplier Selection Problem

4.1.1 Introduction

A supply chain consists of a connected set of activities concerned with planning, coordinating and controlling materials, parts and finished good from supplier to customer. The contribution of the purchasing function to the profitability of the supply chain has assumed greater proportions in recent years; one of the most critical functions of purchasing is selection of suppliers. For most manufacturing firms, the purchasing of raw material and component parts from suppliers constitutes a major expense. Raw material cost accounts for 40–60% of production costs for most US manufacturers. In fact, for the automotive industry, the cost of components and parts from outside suppliers may exceed 50% of sales [1]. For technology firms, purchased materials and services account for 80% of the total production cost. It is vital to the competitiveness of most firms to be able to keep the purchasing cost to a minimum. In today's competitive operating environment it is impossible to successfully produce low-cost, high-quality products without good suppliers. A study carried out by the Aberdeen group [2] found that more than 83% of the organizations engaged in outsourcing achieved significant reduction in purchasing cost, more than 73% achieved reduction in transaction cost and over 60% were able to shrink sourcing cycles.

Supplier selection process is difficult because the criteria for selecting suppliers could be conflicting. Figure 4.1 shows the various factors which could impact the supplier selection process [3]. Supplier selection is a multiple criteria optimization problem that requires trade-off among different qualitative and quantitative factors to find the best set of suppliers. For example, the supplier with the lowest unit price may also have the lowest quality. The problem is also complicated by the fact that several conflicting criteria must be considered in the decision making process.

Most of the times buyers have to choose among a set of suppliers by using some predetermined criteria such as, quality, reliability, technical capability, lead-times etc, even before building long-term relationships. To accomplish these goals, two basic and interrelated decisions must be made by a firm. The firm must decide which suppliers to do business with and how much to order from each supplier. Weber et al. [4] refer to this pair of decisions as the supplier selection problem.

4.1.2 Supplier Selection Process

Figure 4.2 shows the steps in the supplier selection process. The first step is to determine whether to *make or buy* the item. Most organizations buy those parts which are not core to the business or not cost effective if produced in-house. The next step is to define the various criteria for selecting the suppliers. The criteria for selecting a supplier of critical product may not be the same as a supplier of MRO items. Once a decision to buy the item is taken, the most critical step is selecting the right supplier. Once the suppliers are chosen the organization has to negotiate terms of contract and monitor their performance.

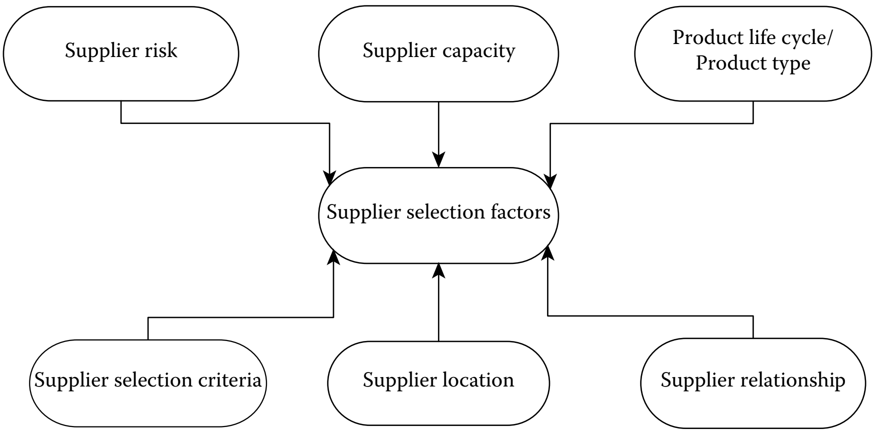


FIGURE 4.1 Supplier selection factors.

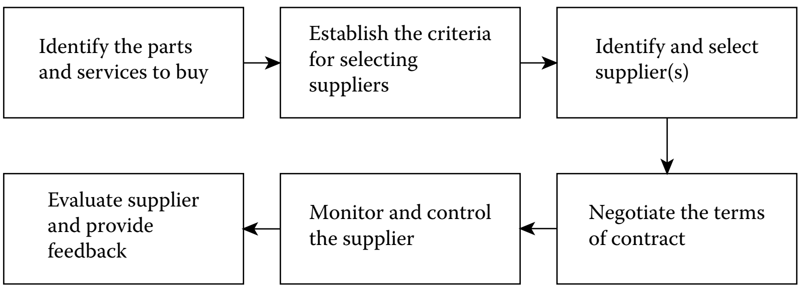


FIGURE 4.2 Supplier selection steps.

Finally, the suppliers have to be constantly evaluated and the feedback should be provided to purchasing and the suppliers.

4.1.3 Supplier Selection in the Military

The first recorded supplier selection model was used by the National Bureau of Standards as a way for awarding procurement contracts in the Department of Defense (DOD) [5–7]. These early studies used a modified form of the transportation problem to model bid evaluations for procurements at the DOD. One method used by the DOD to identify high-quality suppliers and provide them with the incentive to improve their quality is to initially adopt dual sourcing [8]. Based on the performance of the supplier, DOD began to redistribute business toward the better performing suppliers. In the 1970s Pratt and Whitney used to be the sole supplier for the F100 engines. To get higher quality engines from Pratt and Whitney, DOD then started procuring a portion of engines from GE. The government made each companyís share of future business contingent on its productsí quality and performance. The resulting increase in quality due to competition was significant. In testimony before the House Appropriations subcommittee on Defense in 1979, General Lew Allen, Air Force Chief of Staff, explained [9]:

> We are concerned about the motivation and incentive of Pratt and Whitney to correct this engine....[The] best way to insure that we were adequately addressing the problem was to generate some competition.... The approach with General Electric...is an attempt...to develop a true competitive situation within the engine industry.

In recent years, DOD has increased its reliability on commercial suppliers as a way to reduce both purchasing cost as well as procurement time [10]. As of 2006, DOD has a budget of over $400 billion a year, of which about 15% is spent on procurement activities. Any saving resulting from procurement costs can be diverted to R&D or any other value added service [10]. Hence supplier selection plays a key role in the military.

4.1.4 Chapter Overview

In Section 4.2, we review the literature and discuss existing methods that are available for supplier selection. In Section 4.3, we present multiple criteria optimization models for supplier selection and discuss methods for ranking the suppliers. When there are a large number of suppliers, these methods can be used for prequalification or prescreening of the supplier base. The methods are illustrated with an example. In Section 4.4, we present multiple criteria models for supplier order allocation when there are several prequalified suppliers and multiple products to purchase. These models will determine the optimal order quantities from each supplier for the various products under conflicting criteria. In Section 4.4, we also discuss different goal programming (GP) models for solving the multiple criteria supplier selection and order allocation problem. A case study is used to illustrate the various GP approaches. The chapter ends with concluding remarks in Section 4.5.

4.2 Overview of Supplier Selection Methods

4.2.1 Supplier Selection Models

As mentioned earlier, the supplier selection activity plays a key role in cost reduction and is one of the most important functions of the purchasing department. Different mathematical, statistical and game theoretical models have been proposed to solve the problem. Other works provide an overview of supplier selection methods [4,11–13].

De Boer et al. [12] stated that that supplier selection process is made up of several decision making steps as shown in Figure 4.3.

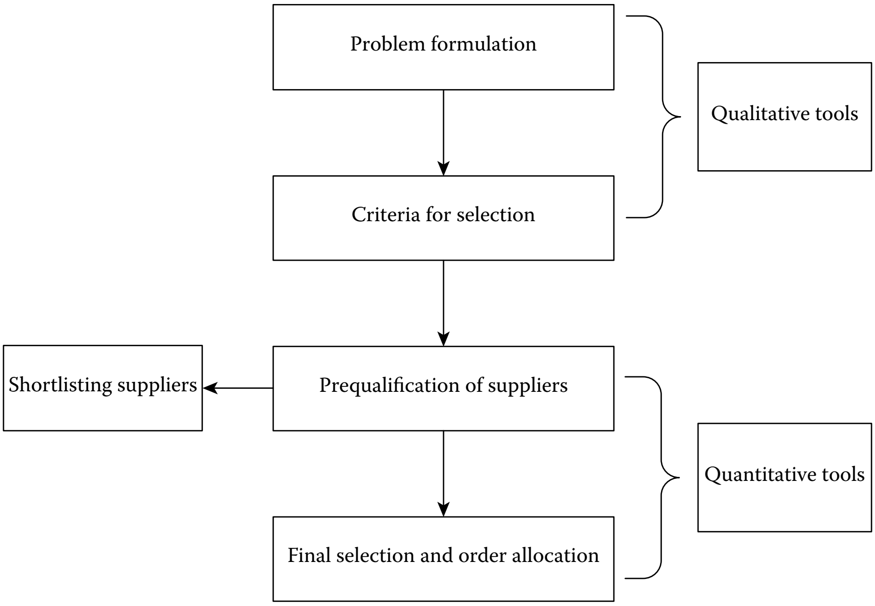


FIGURE 4.3 Supplier selection process.

4.2.1.1 Problem Formulation

Decision support methods for problem formulation are those methods that support the purchasing manager in identifying the need for supplier selection; in this case problem formulation involves determining what the ultimate problem is and why supplier selection is the better choice. According to a survey by Aissaoui et al. [11] and De Boer et al. [12], there are no papers that deal with the problem formulation.

4.2.1.2 Criteria for Selection

Depending on the buying situation, different sets of criteria may have to be employed. Criteria for supplier selection have been studied extensively since the 1960s. Dicksonís study [13] was the earliest to review the supplier selection criteria. He identified 23 selection criteria with varying degrees of importance for the supplier selection process. Dickson's study was based on a survey of purchasing agents and managers. A follow-up study was done by Weber et al. [4]. They reviewed 74 articles published in the 1970s and 1980s and categorized them based on Dickson's 23 selection criteria. In their studies, net price, quality, delivery and service were identified as the key selection criteria. They also found that under just-in-time manufacturing, quality and delivery became more important than net price. Supplier selection criteria may also change over time. Wilson [14] examined studies conducted in 1974, 1982 and 1993 on the relative importance of the selection criteria. She found that quality and service criteria began to dominate price and delivery. She concluded that globalization of the market place and the increasingly competitive atmosphere had contributed to the shift. Table 4.1 shows the top 15 supplier selection criteria as given in Weber et al. [4].

De Boer et al. [12] pointed out that there are only two known techniques for formulating supplier selection criteria. Mandal and Deshmukh [15] proposed interpretive structural modeling (ISM) as technique based on group judgment to identify and summarize relationships between supplier choice criteria through a graphical model. Vokurka et al. [16] developed an expert system for the formulation of the supplier selection criteria. Nominal group technique involving all the stakeholders of the supplier selection decision can also be used to identify the important supplier selection criteria [17].

TABLE 4.1 Criteria For Supplier Selection

Rank	Criteria
1	Net price
2	Delivery
3	Quality
4	Production facilities and capabilities
5	Geographical location
6	Technical capability
7	Management and organization
8	Reputation and position in industry
9	Financial position
10	Performance history
11	Repair service
12	Attitude
13	Packaging ability
14	Operational control
15	Training aids

Source: Weber, C.A., Current J.R., and Benton, W.C. 1991. Supplier selection criteria and methods. *European Journal of Operational Research*, 50, 2.

4.2.1.3 Prequalification of Suppliers

Prequalification is the process of screening suppliers to identify a suitable subset of suppliers. Prequalification reduces the large set of initial suppliers to a smaller set of acceptable suppliers and is more of a sorting process. De Boer et al. [12] have cited many different techniques for prequalification. Some of these techniques are: categorical methods, data envelopment analysis (DEA), cluster analysis, case-based reasoning (CBR) systems and multicriteria decision making methods (MCDM). Several authors have worked on prequalification of suppliers. Weber and Ellram [18], Weber et al. [19] have developed DEA methods for prequalification. Hinkel et al. [20] and Holt [21] used cluster analysis for prequalification and finally Ng and Skitmore [22] developed CBR systems for prequalification. Mendoza et al. [23] developed a three phase multicriteria method to solve a general supplier selection problem. The paper combines analytic hierarchy process (AHP) with GP for both prequalification and final order allocation.

4.2.1.4 Final Selection

Most of the publications in the area of supplier selection have focused on final selection. In the final selection step, the buyer identifies the suppliers to do business with and allocates order quantities among the chosen supplier(s). In reviewing the literature, there are basically two kinds of supplier selection problem, as stated by Ghodsypour and O'Brien [24]:

- *Single sourcing,* which implies that any one of the suppliers can satisfy the buyer's requirements of demand, quality, delivery, etc.
- *Multiple sourcing*, which implies that there are some limitations in suppliers' capacity, quality, etc. and multiple suppliers have to be used.

Specifically, no one supplier can satisfy the buyer's total requirements and the buyer needs to purchase some part of its demand from one supplier, and the other part from another supplier to compensate for the shortage of capacity or low quality of the first supplier. Several methods have been proposed in the literature for single sourcing as well as for multiple sourcing.

Single sourcing models: Single sourcing is a possibility when a relatively small number of parts are procured [11]. Some of the methods used in single sourcing are:

- *Linear weighted point:* The linear weighted point method is the most widely used approach for single sourcing. This approach uses a simple scoring method, which heavily depends on human judgment. Some of the references that discuss this approach include Wind and Robinson [25] and Zenz [26] etc.
- *Cost ratio*: Cost ratio is a more complicated method in which the cost of each criterion is calculated as a percentage of the total purchased value, and a net adjusted cost for each supplier is determined. This approach is complex and needs a lot of financial information. This approach was proposed by Timmerman [27].
- *Analytic hierarchy process (AHP):* AHP, developed by Saaty [28] in the early 1980s, is a good tool to solve multiple criterion problems with finite alternatives. It is a powerful and flexible decision making process to help the decision maker set priorities and make the best decision when both qualitative and quantitative aspects of decision are to be considered. AHP has been extensively used for supplier selection problem [29–34].
- *Total cost of ownership (TCO)*: TCO is a method which looks beyond the price of a product to include other related costs like quality costs, technology costs etc. Finally the business is awarded to the supplier with lowest unit total cost. General Electric Wiring Devices have developed a total cost supplier selection method that takes into account risk factors, business desirable factors and measurable cost factors [35]. TCO approach has also been used by Ellram [36], Degreave [37–39].

Multiple sourcing models: Multiple sourcing can offset the risk of supply disruption. In multiple sourcing, a buyer purchases the same item from more than one supplier. Mathematical programming is the most appropriate method for multiple sourcing decisions. It allows the buyers to consider different constraints while choosing the suppliers and their order allocation. Two types of mathematical programming models are found in the literature, single objective and multiobjective models.

- Single objective models

Moore and Fearon [40] stated that price, quality and delivery are important criteria for supplier selection. They discussed the use of linear programming in the decision making. Gaballa [41] applied mathematical programming to supplier selection in a real case. He used a mixed integer programming to formulate a decision making model for the Australian Post Office. The objective for this approach is to minimize the total discounted price of allocated items to the suppliers. Anthony and Buffa [42] developed a single objective linear programming model to support strategic purchasing scheduling (SPS). The linear model minimized the total cost by considering limitations of purchasing budget, supplier capacities and buyer's demand. Price and storage cost were included in the objective function. The costs of ordering, transportation and inspection were not included in the model. Bender et al. [43] applied single objective programming to develop a commercial computerized model for supplier selection at IBM. They used mixed integer programming, to minimize the sum of purchasing, transportation and inventory costs. Narasimhan and Stoynoff [44] applied a single objective, mixed integer programming model to a large manufacturing firm in the Midwest, to optimize the allocation procurement for a group of suppliers. Turner [45] presented a single objective linear programming model for British Coal. This model minimized the total discounted price by considering the supplier capacity, maximum and minimum order quantities, demand, and regional allocated bounds as constraints. Pan [46] proposed multiple sourcing for improving the reliability of supply for critical materials, in which more than one supplier is used and the demand is split between them. The author used a single objective linear programming model to choose the best suppliers, in which three criteria are considered: price, quality and service. Seshadri et al. [47] developed a probabilistic model to represent the connection between multiple sourcing and its consequences, such as number of bids, the seller's profit and the buyer's price. Benton [48] developed a nonlinear programming model and a heuristic procedure using Lagrangian relaxation for supplier selection under conditions of multiple items, multiple suppliers, resource limitations and quantity discounts. Chaudhry et al. [49] developed linear and mixed integer programming models for supplier selection. In their model price, delivery, quality and quantity discount were included. Papers by Degraeve [37] and Ghodsypour and O'Brien [24] tackled the supplier selection issue in the framework of TCO or total cost of logistics. Jayaraman et al. [50] formulated a mixed integer linear programming model for solving the supplier selection problem with multiple products, buyers and suppliers. Feng et al. [51] presented a stochastic integer-programming model for simultaneous selection of tolerances and suppliers based on the quality loss function and process capability index.

- Multiobjective models

Among the different multicriteria approaches for supplier selection, GP is the most commonly used method. Buffa and Jackson [52] presented a multicriteria linear GP model. In this model two sets of factors are considered: supplier attributes, which include quality, price, service experience, early, late and on-time deliveries, and the buying firm's specifications, including material requirement and safety stock. Sharma et al. [53] proposed a GP formulation for attaining goals pertaining to price, quality and lead-time under demand and budget constraints. Liu et al. [54] developed a decision support system by integrating AHP with linear programming. Weber et al. [19] used multiobjective linear programming for supplier selection to systematically analyze the trade-off between conflicting factors. In this model aggregate price, quality and late delivery were considered as goals. Karpak et al. [55] used visual interactive GP for supplier selection process. The objective was to identify and allocate order quantities among suppliers while minimizing product acquisition cost and maximizing quality and reliability. Bhutta and Huq [56] illustrated and compared the technique of TCO and AHP in supplier selection process. Wadhwa and Ravindran [1] formulated a supplier selection problem with price, lead-time and quality as three conflicting objectives. The suppliers offered quantity discounts and the model was solved using GP, compromise programming and weighted objective methods.

Table 4.2 gives a brief summary of some of the other papers not reviewed in this section but have been published since 2000 dealing with supplier selection models and methods.

4.3 Multicriteria Ranking Methods for Supplier Selection

Many organizations have a large pool of suppliers to select from. The supplier selection problem can be solved in two phases. The first phase reduces the large number of candidate suppliers to a manageable size. In Phase two, a multiple criteria optimization model is used to allocate order quantities among the short listed suppliers. Figure 4.4 shows the steps involved in the two phase supplier selection model.

4.3.1 Prequalification of Suppliers

Prequalification is defined as the process of reducing the large set of suppliers to a smaller manageable number by ranking the suppliers under a pre defined set of criteria. The primary benefits of prequalification of suppliers are [21]:

1. The possibility of rejecting good suppliers at an early stage is reduced.
2. Resource commitment of the buyer toward purchasing process is optimized.
3. With the application of preselected criteria, the prequalification process is rationalized.

In this section, we present multiple criteria ranking approaches for the supplier selection problem, namely, the prequalification of suppliers.

TABLE 4.2 Supplier Selection Methods

Author	Method	Brief Description
Yang [57]	Multicriteria Math Model	Developed a five step multi-criteria strategic supplier selection model incorporating the supply risk.
Xia and Wu [58]	AHP with multi objective mathematical programming	Formulated a multicriteria supplier selection problem with supplier price breaks. Incorporates AHP to calculate criteria weights to be used in the model.
Saen [59]	DEA (imprecise DEA)	Used a modified version of DEA to include both qualitative and quantitative data in supplier selection
Chen et al. [60]	Fuzzy TOPSIS	Used fuzzy numbers to handle linguistic judgments and applied TOPSIS MCDM technique to address the supplier selection problem.
Haq and Kannan [61]	Fuzzy AHP, genetic algorithm	Addressed the supplier selection using fuzzy AHP. Formulates a multi echelon supply chain configuration problem and solves it using GA.
Cachon and Zhang [62]	Game theory	Formulated a game theoretic supplier selection model under information asymmetry.
Ding et al. [63]	Simulation, genetic Algorithm	Proposed a hybrid method where GA is used to search for optimal supplier portfolios and simulation is used to estimate key supplier performance parameters.
Piramuthu [64]	Agent based	Proposed an agent based model with a learning component.
Deng and Elmaghraby [8]	Tournament, game theory	Proposed a tournament type model where suppliers compete against each other.
Liu and Wu [65]	AHP and DEA	Developed an integrated method by combining AHP and DEA, applied to the supplier selection problem.
Valluri and Croson [66]	Reinforcement learning (RL)	Proposed a RL approach to supplier selection with two separate selection policies.
Emerson and Piramuthu [67]	Agent based	Proposed an agent based manufacturing supplier selection model with a learning component.
Agrell et al. [68]	Game theory	Applied a game theoretic model to select suppliers in a telecommunication firm.
Chan [69]	AHP	Developed an interactive technique to facilitate data collection prior to AHP implementation in supplier selection problems.
Choy et al. [70]	Neural networks	Used neural networks to benchmark suppliers.

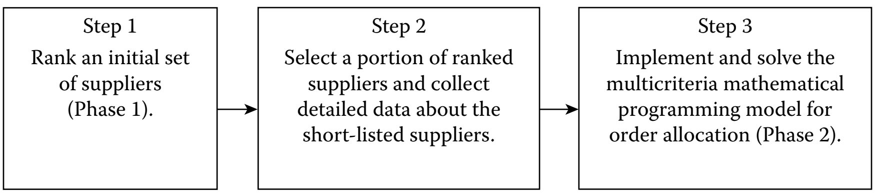


FIGURE 4.4 Two phase supplier selection model.

In the prequalification process (Phase 1), readily available qualitative and quantitative data are collected for the various suppliers. This data can be obtained from trade journals, internet and past transactions to name a few. Once this data is gathered, these suppliers are evaluated using multiple criteria ranking methods. The decision maker then selects a portion of the suppliers for extensive evaluation in Phase 2.

The first step in prequalification is defining the selection criteria. We have used the following 14 prequalification criteria as an illustration. The prequalification criteria have been split into various subcategories such as organizational criteria, experience criteria etc. The various prequalification criteria are described below:

- Organizational criteria:
  - Size of company (C1): Size of the company can be either its number of employees, or its market capitalization.
  - Age of company (C2): Age of the company is the number of years that the company has been in business.
  - R&D activities (C3): Investment in research and development.
- Experience criteria:
  - Project type (C4): Specific types of projects completed in the past.
  - Project size (C5): Specific sizes of projects completed in the past.
- Delivery criteria:
  - Accuracy (C6): Meeting the promised delivery time.
  - Capacity (C7): Capacity of the supplier to fulfill orders.
  - Lead-time (C8): Supplier's promised delivery lead-time.
- Quality criteria:
  - Responsiveness (C9): If there is an issue concerning quality, how fast the supplier reacts to correct the problem.
  - Defective rate (C10): Rate of defective items among orders shipped.
- Cost criteria:
  - Order change and cancellation charges (C11): Fee associated with modifying or changing orders after they have been placed.
  - Unit cost: Price per item (C12).
- Miscellaneous criteria:
  - Labor relations (C13): Number of strikes or any other labor problems encountered in the past.
  - Procedural compliances (C14): Conformance to national/international standards (e.g. ISO 9000).

For the sake of illustration, we assume there are 20 suppliers during prequalification. The 14 supplier criteria values for the initial set of 20 suppliers are given in Table 4.3. For all the supplier criteria, larger values are preferred. Next, we discuss several multiple criteria ranking methods for short listing the suppliers. Each method has advantages and limitations. The methods that we discuss here are:

1. L_p metric method
2. Rating method
3. Borda count
4. AHP
5. Cluster analysis

For a more detailed discussion of multicriteria ranking methods, the reader is referred to Ravindran [71].

4.3.2 Use of L_p Metric for Ranking Suppliers

Mathematically, the L_p metric represents the distance between two vectors **x** and **y**, where $\mathbf{x}, \mathbf{y} \in R^n$, and is given by:

$$\|\mathbf{x}-\mathbf{y}\|_p = \left[\sum_{j=1}^{n} | x_j - y_j |^p\right]^{1/p} \tag{4.1}$$

One of the most commonly used L_p metrics is the L_2 metric ($p=2$), which measures the Euclidean distance between two vectors. The ranking of suppliers is done by calculating the L_p metric between the ideal solution (H) and each vector representing the supplier's ratings for the criteria. The ideal solution represents the best values possible for each criterion from the initial list of suppliers. Since no supplier will have the best values for all criteria (e.g. a supplier with minimum cost may have poor quality and delivery time), the ideal solution is an artificial target and cannot be achieved. The L_p metric approach computes the distance of each supplier's attributes from the ideal solution and ranks the supplier's based on that distance (smaller the better). We shall illustrate the steps of the L_2 metric method using the supplier data in Table 4.3.

4.3.2.1 Steps of the L_2 Metric Method

Step 1: Determine the ideal solution. The ideal values (H) for the 14 criteria of Table 4.3 are given in Table 4.4.

Step 2: Use the L_2 metric to measure the closeness of supplier to the ideal values. The L_2 metric for supplier k is given by

$$L_2(k)=\sqrt{\sum_{j=1}^{n}(H_j-Y_{jk})^2} \tag{4.2}$$

where H_j is the ideal value for criterion j and Y_{jk} is the j^{th} criterion value for supplier k.

Step 3: Rank the suppliers using the L_2 metric. The supplier with the smallest L_2 value is ranked first, followed by the next smallest L_2 value etc. Table 4.5 gives the L_2 distance from the ideal for each supplier and their resultant rankings.

TABLE 4.3 Supplier Criteria Values

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
S1	0.75	1	0.46	1	0.92	0.9	1	0	0.13	0.18	0.18	0.01	0.26	0.79
S2	0.22	0	0.33	1	0.94	0.35	0.9	0.13	0.02	0	0.38	0.95	0.88	0.72
S3	0.53	0	0.74	0	0.03	0.89	0.1	0.12	0	0.3	0.66	0.08	0.86	0.22
S4	0.28	1	0.8	0	0.54	0.75	0.85	1	1	0.87	0.33	0.5	0.78	0.12
S5	0.3	0	0.79	1	0.6	0.49	0.8	0.15	0.97	0.79	0.83	0.13	0.46	0.15
S6	0.5	1	0.27	0	0.43	0.52	0.12	0	0	0.25	0.9	0.07	0.26	0
S7	0.25	1	0.6	1	0.1	0.18	0	0.13	1	0.85	0.51	0.59	0.12	1
S8	0.76	1	0.68	1	0.55	0.87	0	0.14	0	1	0.98	0.19	0.86	0.99
S9	0.25	1	0.5	1	0.26	0.92	0.94	0.03	0.15	1	0.7	0.41	0.95	1
S10	0.16	1	0.7	0	0.46	0.62	0.9	0	0.03	0	0.3	0.68	0.61	1
S11	0.31	0	0.3	0	0.09	0.73	1	1	1	0	0.87	0.3	0.98	0
S12	0.34	1	0.39	1	0.75	0.94	0.78	0.3	0	0.85	0.94	0.61	0.46	0.3
S13	0.08	0	0.27	0	0.14	0.42	1	0.91	0	0.82	0.45	0.42	0.81	1
S14	0.62	1	0.02	1	0.15	0.97	0.15	0.01	0.18	0.92	0.55	0.23	0.12	0.97
S15	0.49	0	0.98	0	0.52	0.68	0	0.24	0.06	0	0.52	0.84	0.05	0.76
S16	0.1	1	0.32	1	0.67	0.21	1	0.85	0.16	0.29	0.49	0.41	0.29	0.27
S17	0.08	0	0.19	1	0.24	0.87	0	0.72	0.26	1	0.84	0.99	0.64	0.04
S18	0.86	0	0.28	1	0.95	0.08	1	0.12	0.2	0	0.4	0.76	0.66	1
S19	0.72	0	0.88	0	0.15	0.93	0.97	1	1	1	0.75	0.64	0.26	1
S20	0.15	1	0.92	1	0.77	0.63	0	0	0.3	0.22	0.22	0.94	0.93	0.26

TABLE 4.4 Ideal Values (*H*)

Criteria	Ideal Value	Criteria	Ideal Value
C1	0.86	C8	1
C2	1	C9	1
C3	0.98	C10	1
C4	1	C11	0.18
C5	0.95	C12	0.99
C6	0.08	C13	0.05
C7	1	C14	1

TABLE 4.5 Supplier Ranking Using L_2 Metric

Supplier	L_2 value	Rank	Supplier	L_2 value	Rank
Supplier1	2.105	7	Supplier11	2.782	18
Supplier2	2.332	11	Supplier12	2.083	5
Supplier3	3.011	20	Supplier13	2.429	15
Supplier4	1.896	3	Supplier14	2.347	13
Supplier5	2.121	8	Supplier15	2.517	16
Supplier6	2.800	19	Supplier16	1.834	2
Supplier7	1.817	1	Supplier17	2.586	17
Supplier8	2.357	4	Supplier18	2.092	6
Supplier9	2.206	9	Supplier19	1.970	4
Supplier10	2.339	12	Supplier20	2.295	10

4.3.3 Rating (Scoring) Method

Rating is one of the simplest and most widely used ranking methods under conflicting criteria. First, an appropriate rating scale is agreed to (e.g. from 1 to 10, where 10 is the most important and 1 is the least important selection criteria). The scale should be clearly understood to be used properly. Next, using the selected scale, the decision maker (DM) provides a rating r_j for each criterion, C_j. The same rating can be given to more than one criterion. The ratings are then normalized to determine the weights of the criteria *j*. Assuming *n* criteria:

$$W_j = \frac{r_j}{\sum_{j=1}^{j=n} r_j} \quad \text{for } j = 1, 2, \ldots, n \tag{4.3}$$

Note

$$\sum_{j=1}^{n} w_j = 1.$$

Next, a weighted score of the attributes is calculated for each supplier as follows:

$$S_k = \sum_{j=1}^{n} W_j f_{jk} \quad \text{for } k = 1 \ldots K$$

where f_{jk}'s are the criteria values for supplier *k*. The supplier's are then ranked based on their scores. The supplier with the highest score is ranked first. Rating method requires relatively little cognitive burden on the DM.

TABLE 4.6 Initial Supplier Ranking Using Rating Method

Criterion	Rating	Weight	Criterion	Rating	Weight
C1	6	0.073	C8	1	0.012
C2	7	0.085	C9	8	0.098
C3	5	0.061	C10	7	0.085
C4	9	0.110	C11	6	0.073
C5	10	0.122	C12	7	0.085
C6	2	0.024	C13	4	0.049
C7	3	0.037	C14	7	0.085

TABLE 4.7 Supplier Ranking Using Rating Method

Supplier	Total Score	Rank	Supplier	Total Score	Rank
Supplier1	0.571	8	Supplier11	0.353	18
Supplier2	0.518	12	Supplier12	0.631	3
Supplier3	0.255	20	Supplier13	0.360	17
Supplier4	0.568	9	Supplier14	0.532	11
Supplier5	0.565	10	Supplier15	0.361	16
Supplier6	0.312	19	Supplier16	0.503	13
Supplier7	0.616	4	Supplier17	0.477	14
Supplier8	0.681	1	Supplier18	0.577	7
Supplier9	0.649	2	Supplier19	0.584	5
Supplier10	0.441	15	Supplier20	0.583	6

Table 4.6 illustrates the ratings and the corresponding weights for the 14 criteria.

Table 4.7 shows the final scores for different suppliers and their rankings using the rating method.

4.3.4 Borda Count

This method is named after Jean Charles de Borda, an 18th century French physicist. The method is as follows:

- The n criteria are ranked 1 (most important) to n (least important)
  - Criterion ranked 1 gets n points, 2nd rank gets n–1 points, and the last place criterion gets 1 point.
- Weights for the criteria are calculated as follows:
  - Criterion ranked $1 = n/S$
  - Criterion ranked $2 = (n-1)/S$
  - Last criterion $= 1/S$

where S is the sum of all the points $= n(n+1)/2$.

Table 4.8 shows the calculations of criteria weights using the Borda count method. For example, criterion 3 is ranked first among the 14 criteria and gets 14 points. Criterion 11 is ranked last and gets 1 point. Thus, the weight for criterion 3 = 14/105 = 0.133. Using these criteria weights, the supplier scores are calculated as before for ranking as shown in Table 4.9. Note that when there are many criteria, it would be difficult for a DM to rank order them precisely. In practice, pair wise comparison of criteria is used to facilitate the criteria ranking required by the Borda count. Here, the DM is asked to give the relative importance between two criteria C_i and C_j, whether C_i is preferred to C_j, C_j is preferred to C_i or both are equally important. When there are n criteria, the DM has to respond to $n(n-1)/2$ pair wise comparisons. Based on the DM's response, it is easy to determine the ranking of the criteria, from which the criteria weights can be computed using the Borda count.

TABLE 4.8 Criterion Rating Using Borda Count

Criterion	Ranking Points	Weight	Criterion	Ranking Points	Weight
C1	9	0.086	C8	13	0.124
C2	7	0.067	C9	2	0.019
C3	14	0.133	C10	8	0.076
C4	6	0.057	C11	1	0.010
C5	5	0.048	C12	12	0.114
C6	10	0.095	C13	4	0.038
C7	11	0.105	C14	3	0.029

TABLE 4.9 Final supplier Ranking Using Borda Count

Supplier	Total Score	Rank	Supplier	Total Score	Rank
Supplier1	0.536	6	Supplier11	0.468	15
Supplier2	0.476	14	Supplier12	0.619	3
Supplier3	0.333	19	Supplier13	0.481	13
Supplier4	0.686	2	Supplier14	0.434	17
Supplier5	0.489	11	Supplier15	0.417	18
Supplier6	0.274	20	Supplier16	0.533	7
Supplier7	0.452	16	Supplier17	0.500	9
Supplier8	0.575	5	Supplier18	0.489	10
Supplier9	0.612	4	Supplier19	0.714	1
Supplier10	0.482	12	Supplier20	0.530	8

4.3.5 AHP

The AHP, developed by Saaty [28], is a MCDM method for ranking alternatives. Using AHP, the decision maker can assess not only quantitative but also various intangible factors such as financial stability, feeling of trust etc in the supplier selection process. The buyer establishes a set of evaluation criteria and AHP uses these criteria to rank the different suppliers. AHP can enable DM to represent the interaction of multiple factors in complex and un-structured situations.

4.3.5.1 Basic Principles of AHP

- Design a hierarchy: Top vertex is the main objective and bottom vertices are the alternatives. Intermediate vertices are criteria/subcriteria (which are more and more aggregated as you go up in the hierarchy).
- At each level of the hierarchy, a paired comparison of the vertices criteria/subcriteria is performed from the point of view of their "contribution (weights)" to each of the higher-level vertices to which they are linked.
- Uses both rating method and comparison method. A numerical scale 1–9 (1-equal importance; 9-most important).
- Uses pair wise comparison of alternatives with respect to each criterion (subcriterion) and gets a numerical score for each alternative on every criterion (subcriterion).
- Compute total weighted score for each alternative and rank the alternatives accordingly.

To design the hierarchy in our example, the 14 criteria are grouped into six major criteria and several subcriteria as shown in Figure 4.5.

4.3.5.2 Steps of the AHP Model

Step 1: In the first step, carry out a pair wise comparison of criteria using the 1–9 degree of importance scale shown in Table 4.10.

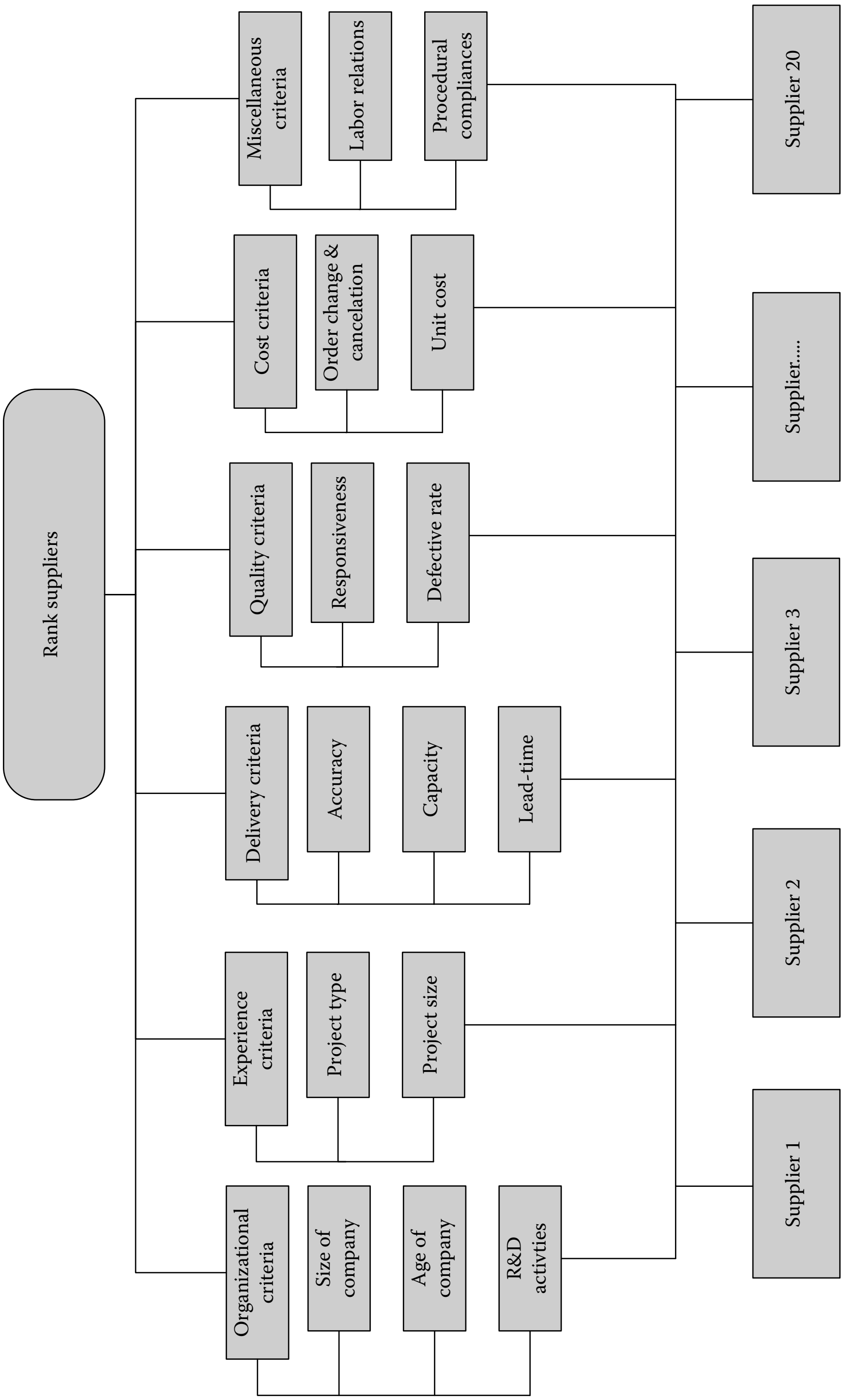


FIGURE 4.5 AHP hierarchy for the example problem.

TABLE 4.10 Degree of Importance Scale in AHP

Degree of Importance	Definition
1	Equal importance
3	Weak importance of one over other
5	Essential or strong importance
7	Demonstrated importance
9	Absolute importance
2, 4, 6, 8	Intermediate values between two adjacent judgments

TABLE 4.11 Pair Wise Comparison of Criteria

	Organizational	Experience	Delivery	Quality	Cost	Miscellaneous
Organizational	1	0.2	0.143	0.33	0.33	1
Experience	5	1	0.5	2	2	5
Delivery	7	2	1	5	4	7
Quality	3	0.5	0.2	1	1	3
Cost	3	0.5	0.25	1	1	3
Miscellaneous	1	0.2	0.143	0.33	0.33	1

TABLE 4.12 Final Criteria Weights Using AHP

Criteria	Weight
Organizational	0.047
Experience	0.231
Delivery	0.430
Quality	0.120
Cost	0.124
Miscellaneous	0.047

If there are n criteria to evaluate, then the pair wise comparison matrix for the criteria is given by $A_{(N\times N)}=[a_{ij}]$, where a_{ij} represents the relative importance of criterion i with respect to criterion j. Set $a_{ii}=1$ and $a_{ji}=1/a_{ij}$. The pair wise comparisons with the degree of importance for the six major criteria is shown in Table 4.11.

Step 2: Compute the normalized weights for the main criteria. We obtain the weights using L_1 norm. The two step process for calculating the weights is as follows:

- Normalize each column of $A_{(N\times N)}$ using L_1 norm.

$$r_{ij}=\frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}} \tag{4.4}$$

- Average the normalized values across each row.

$$w_i=\frac{\sum_{j=1}^{n} r_{ij}}{n} \tag{4.5}$$

Table 4.12 shows the results obtained as a result of Step 2.

Step 3: In the third step we check for consistency of the pair wise comparison matrix using the eigen value theory as follows [28].

1. Using the pair wise comparison matrix A (Table 4.11) and the weights W (Table 4.12) compute AW. Let the vector $X=(X_1, X_2, X_3, \ldots, X_n)$ denote the values of AW.
2. Compute

$$\lambda_{\max} = \text{Average}\left[\frac{X_1}{W_1}, \frac{X_2}{W_2}, \frac{X_3}{W_3} \cdots\cdots\cdots \frac{X_n}{W_n}\right] \tag{4.6}$$

3. Consistency index (CI) is given by

$$\text{CI} = \frac{\lambda_{\max} - n}{n-1} \tag{4.7}$$

Saaty [28] generated a number of random positive reciprocal matrices with $a_{ij} \in (1, 9)$ for different sizes and computed their average CI values, denoted by RI:

N	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.52	0.89	1.11	1.25	1.35	1.4	1.45	1.49

He defines the consistency ratio (CR) as CR=CI/RI. If CR<0.15, then accept the pair wise comparison matrix as consistent.

Using the above steps, CR is found to be 0.004 for our example problem. Since the CR is less than 0.15, the response can be assumed to be consistent.

Step 4: In the next step, we compute the relative importance of the subcriteria in the same way as done for the main criteria. Step 2 and Step 3 are carried out for every pair of subcriteria with respect to their main criterion. The final weights of the subcriteria are the product of the weights along the corresponding branch. Table 4.13 shows the final weights of the various criteria and subcriteria.

Step 5: Repeat Steps 1, 2 and 3 and obtain,

TABLE 4.13 AHP Sub-criteria Weights for the Example Problem

Criteria (Criteria Weight)	Subcriteria	Subcriteria Weight	Global Weight (Criteria Weight * Subcriteria Weight)
Organizational (0.047)	Size of company	0.143	0.006
	Age of company	0.429	0.020
	R&D activities	0.429	0.020
Experience (0.231)	Project type	0.875	0.202
	Project size	0.125	0.028
Delivery (0.430)	Accuracy	0.714	0.307
	Capacity	0.143	0.061
	Lead-time	0.143	0.061
Quality (0.120)	Responsiveness	0.833	0.099
	Defective-rate	0.167	0.020
Cost (0.124)	Order change	0.833	0.103
	Unit cost	0.167	0.020
Miscellaneous (0.047)	Labor relations	0.125	0.005
	Procedural compliances	0.875	0.041

TABLE 4.14 Supplier Ranking Using AHP

Supplier	Total Score	Rank
Supplier 1	0.119	20
Supplier 2	0.247	14
Supplier 3	0.325	12
Supplier 4	0.191	18
Supplier 5	0.210	16
Supplier 6	0.120	19
Supplier 7	0.249	13
Supplier 8	0.328	11
Supplier 9	0.192	17
Supplier 10	0.212	15
Supplier 11	0.427	7
Supplier 12	0.661	1
Supplier 13	0.431	6
Supplier 14	0.524	5
Supplier 15	0.431	8
Supplier 16	0.539	4
Supplier 17	0.637	2
Supplier 18	0.421	9
Supplier 19	0.412	10
Supplier 20	0.543	3

a. Pair-wise comparison of alternatives with respect to each criterion using the ratio scale (1–9).
b. Normalized scores of all alternatives with respect to each criterion. Here, an $(m \times m)$ matrix S is obtained, where S_{ij} = normalized score for alternative i with respect to criterion j and m is the number of alternatives.

Step 6: Compute the total score (TS) for each alternative as follows $TS_{(m\times 1)} = S_{(m\times n)}\, W_{(n\times 1)}$, where W is the weight vector obtained after Step 3 and Step 4. Using the total scores, the alternatives are ranked. The total scores obtained by AHP for the supplier selection example is given in Table 4.14. Note: There is commercially available software for AHP called Expert Choice. Interested readers can refer to http://www.expertchoice.com for additional information.

4.3.6 Cluster Analysis

Clustering analysis (CA) is a statistical technique particularly suited to grouping of data. It is gaining wide acceptance in many different fields of research such as data mining, marketing, operations research and bioinformatics. CA is used when it is believed that the sample units come from an unknown population. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait. CA develops subsets of the raw data such that each subset contains member of like nature (similar supplier characteristics) and that difference between different subsets is as pronounced as possible.

There are two types of clustering algorithms [72], namely:

- *Hierarchical:* Algorithms which employ hierarchical clustering find successive clusters using previously established clusters. Hierarchical algorithms can be further classified as *agglomerative* or *divisive*. Agglomerative algorithms begin with each member as a separate cluster and merge them into successively larger clusters. On the other hand, divisive algorithms begin with the whole set as one cluster and proceed to divide it into successively smaller clusters. Agglomerative method is the most common hierarchical method.

- *Partitional:* In partitional clustering, the algorithm finds all the clusters at once. An example of partitional methods is K-means clustering.

4.3.6.1 Procedure for CA

Clustering process begins with formulating the problem and concludes with carrying out analysis to verify the accuracy and appropriateness of the method. The clustering process has the following steps:

1. Formulate the problem and identify the selection criteria.
2. Decide on the number of clusters.
3. Select a clustering procedure.
4. Plot the dendrogram (A dendrogram is a tree diagram used to illustrate the output of clustering analysis) and carry out analysis to compare the means across various clusters.

Let us illustrate the cluster analysis using our supplier selection example.

Step 1: In the first step, every supplier is rated on a scale of 0–1 for each attribute as shown in Table 4.3.

Step 2: In this step, we need to decide on the number of clusters. We want the initial list of suppliers to be split into two categories, good suppliers and bad suppliers; hence the number of clusters is two.

Step 3: Next, we apply both hierarchical and partitional clustering methods to supplier data. We choose the method which has the highest R-sq value pooled over all the 14 attributes. A summary table showing the pooled R-sq value is shown in Table 4.15.

The R-sq value for k-means is the highest among different methods; hence k-means is chosen for clustering. There are several other methods available for determining the goodness of fit [72].

Step 4: In this step, we apply the k-means clustering to the given data. Graphical comparison of the two clusters is shown in the Figure 4.6. Figure 4.6 helps to identify the cluster which has suppliers of higher mean scores. Cluster 1 has six members and cluster 2 has 14 members. From Figure 4.6, it can be seen that suppliers of cluster 1 have a better mean scores than suppliers in cluster 2 (on most criteria); hence 6 suppliers of cluster 1 are chosen as short-listed suppliers.

4.3.7 Comparison of Ranking Methods

Different ranking methods can provide different solutions resulting in rank reversals. In extensive empirical studies with human subjects, it has been found [73,74] that Borda count (with pair wise comparison of criteria) rankings are generally in line with AHP rankings. Given the increased cognitive burden and expensive calculations required for AHP, Borda count might be selected as an appropriate method for supplier rankings. Even though the rating method is easy to use, it could lead to several ties in the final rankings, there by making the results less useful.

TABLE 4.15 Pooled R-Sq Value Different Cluster Methods

Method	Pooled R-Sq Over all the Variables
Single linkage	0.05
Wards linkage	0.1301
Centroid linkage	0.06
Complete linkage	0.1307
Average linkage	0.1309
K-means	0.135

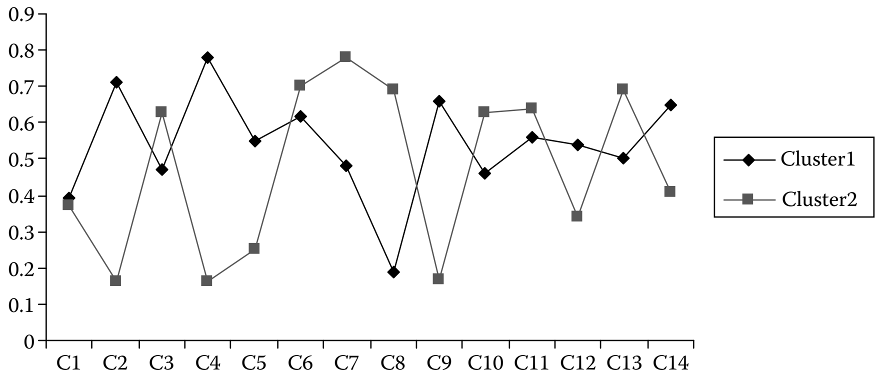


FIGURE 4.6 Comparison of mean across clusters.

4.3.8 Group Decision Making

Most purchasing decisions, including the ranking and selection of suppliers, involve the participation of multiple DMs and the ultimate decision is based on the aggregation of DM's individual judgments to arrive at a group decision. The rating method, Borda count and AHP discussed in this section can be extended to group decision making as described below:

1. *Rating method:* Ratings of each DM for every criterion is averaged. The average ratings are then normalized to obtain the group criteria weights.
2. *Borda count:* Points are assigned based on the number of DMs that assign a particular rank for a criterion. These points are then totaled for each criterion and normalized to get criteria weights. (This is similar to how the college polls are done to get the top 25 football or basketball teams.)
3. *AHP:* There are two methods to get the group rankings using AHP.

   a. Method 1: Strength of preference scores assigned by individual DMs are aggregated using geometric means and then used in the AHP calculations.
   b. Method 2: First, all the alternatives are ranked by each DM using AHP. The individual rankings are then aggregated to a group ranking using Borda count.

4.4 Multiobjective Supplier Allocation Model

As a result of prequalification, the large number of initial suppliers is reduced to a manageable size. In the second phase of the supplier selection, detailed quantitative data such as price, capacity, quality etc. are collected on the shortlisted suppliers and are used in a multiobjective framework for the actual order allocation. We consider multiple buyers, multiple products and multiple suppliers with volume discounts. The scenario of multiple buyers is possible in case of a central purchasing department, where different divisions of an organization buy through one purchasing department. Here, the number of buyers will be equal to the number of divisions buying through the central purchasing. In all other cases, the number of buyers is equal to one. We consider the least restrictive case where any of the buyers can acquire one or more products from any suppliers, namely, a multiple sourcing model.

In this phase of the supplier selection process, an organization will make the following decisions.

- To choose the most favorable suppliers who would meet its supplier selection criteria for the various components.
- To order optimal quantities from the chosen most favorable suppliers to meet its production plan or demand.

The mathematical model for the order allocation problem is discussed next.

4.4.1 Notations used in the Model

Model indices:

- I Set of products to be purchased.
- J Set of buyers who procure multiple units in order to fulfill some demand.
- K Potential set of suppliers.
- M Set of incremental price breaks for volume discounts.

Model parameters:

- p_{ikm} Cost of acquiring one unit of product i from supplier k at price level m.
- b_{ikm} Quantity at which incremental price breaks occurs for product i by supplier k.
- F_k Fixed ordering cost associated with supplier k.
- d_{ij} Demand of product i for buyer j.
- l_{ijk} Lead time of supplier k to produce and supply product i to buyer j. The lead time of different buyers could be different because of geographical distances.
- q_{ik} Quality that supplier k maintains for product i, which is measured as percent of defects.
- CAP_{ik} Production capacity for supplier k for product i.
- N Maximum number of suppliers that can be selected.

Decision variables in the model:

- X_{ijkm} Number of units of product i supplied by supplier k to buyer j at price level m.
- Z_k Denotes if a particular supplier is chosen or not. This is a binary variable which takes a value 1 if a supplier is chosen to supply any product and is zero, if the supplier is not chosen at all.
- Y_{ijkm} This is binary variable which taken on a value 1 if price level m is used, 0 otherwise.

4.4.2 Mathematical Formulation of the Order Allocation Problem

The conflicting objectives used in the model are simultaneous minimization of price, lead-time and rejects. It is relatively easy to include other objectives also. The mathematical form for these objectives is:

1. Price (z_1): Total cost of purchasing has two components; fixed and the variable cost.
   Total variable cost: The total variable cost is the cost of buying every additional unit from the suppliers and is given by:

$$\sum_i \sum_j \sum_k \sum_m p_{ikm} \cdot X_{ijkm} \tag{4.8}$$

   Fixed cost: If a supplier k is used then there is a fixed cost associated with it, which is given by:

$$\sum_k F_k \cdot Z_k \tag{4.9}$$

   Hence the total purchasing cost is,

$$\sum_i \sum_j \sum_k \sum_m p_{ikm} \cdot X_{ijkm} + \sum_k F_k \cdot Z_k \tag{4.10}$$

2. Lead-time (z_2):

$$\sum_i \sum_j \sum_k \sum_m l_{ijk} \cdot X_{ijkm} \tag{4.11}$$

The product of lead-time of each product and quantity supplied is summed over all the products, buyers and suppliers and should be minimized.

3. Quality (z_3):

$$\sum_i \sum_j \sum_k \sum_m q_{ik} \cdot X_{ijkm} \tag{4.12}$$

The product of rejects and quantity supplied is summed over all the products, buyers and suppliers and should be minimized. Quality in our case is measured in terms of percentage of rejects.

The constraints in the model are as follows:

1. *Capacity constraint:* Each supplier k has a maximum capacity for product i, CAP_{ik}. Total order placed with this supplier must be less than or equal to the maximum capacity. Hence the capacity constraint is given by:

$$\sum_i \sum_j \sum_m X_{ijkm} \leq (\text{CAP}_{ik})\text{Z}_\text{k} \quad \forall k \tag{4.13}$$

The binary variable on the right hand side of the constraint implies that a supplier cannot supply any products if not chosen, i.e. if Z_k is 0.

2. *Demand constraint:* The demand of buyer j for product i has to be satisfied using a combination of the suppliers. The demand constraint is given by:

$$\sum_k \sum_m X_{ijkm} = d_{ij} \quad \forall i, j \tag{4.14}$$

3. *Maximum number of suppliers:* The maximum number of suppliers chosen must be less than or equal to the specified number. Hence this constraint takes the following form:

$$\sum_k Z_k \leq N \tag{4.15}$$

4. *Linearizing constraints:* In the presence of incremental price discounts, objective function is nonlinear. The following set of constraints are used to linearize it:

$$X_{ijkm} \leq (b_{ikm} - b_{ikm-1}) * Y_{ijkm} \quad \forall\, i, j, k,\ 1 \leq m \leq m_k \tag{4.16}$$

$$X_{ijkm} \geq (b_{ikm} - b_{ikm-1}) * Y_{ijkm+1} \quad \forall\, i, j, k,\ 1 \leq m \leq m_k - 1 \tag{4.17}$$

$0 = b_{i,\text{k},0} < b_{i,k,1} < \ldots\ldots\ldots < b_{i,k,m_k}$ is the sequence of quantities at which price break occurs. p_{ikm} is the unit price of ordering X_{ijkm} units from supplier k at level m, if $b_{i,k,m-1} < X_{ijkm} \leq b_{i,k,m}$ $(1 \leq m \leq m_k)$.

Constraint 4.16 and Constraint 4.17 force quantities in the discount range for a supplier to be incremental. Because the "quantity" is incremental, if the order quantity lies in discount interval m, namely, $Y_{ijkm} = 1$, then the quantities in interval 1 to m–1, should be at the maximum of those ranges. Constraint 4.16 also assures that a quantity in any range is no greater than the width of the range.

5. *Nonnegativity and binary constraint*: $X_{ijkm} \geq 0$, Z_k, $Y_{ijkm} \in (0,1)$ (4.18)

4.4.3 GP Methodology [71]

The mathematical model formulated in Section 4.4.2 is a multiple objective integer linear programming problem. One way to treat multiple criteria is to select one criterion as primary and the other criteria as secondary. The primary criterion is then used as the optimization objective function, while the secondary criteria are assigned acceptable minimum and maximum values and are treated as problem constraints. However, if careful considerations were not given while selecting the acceptable levels, a feasible solution that satisfies all the constraints may not exist. This problem is overcome by *goal programming*, which has become a popular practical approach for solving multiple criteria optimization problems.

GP falls under the class of methods that use completely pre specified preferences of the decision maker in solving the multicriteria mathematical programming problems. In GP, all the objectives are assigned target levels for achievement and a relative priority on achieving those levels. GP treats these targets as *goals to aspire for* and not as absolute constraints. It then attempts to find an optimal solution that comes as "close as possible" to the targets in the order of specified priorities. In this section, we shall discuss how to formulate GP models and their solution methods.

Before we discuss the formulation of GP problems, we discuss the difference between the terms *real constraints* and *goal constraints* (or simply *goals*) as used in GP models. The real constraints are absolute restrictions on the decision variables, while the goals are conditions one would like to achieve but are not mandatory. For instance a real constraint given by

$$x_1+x_2=3 \tag{4.19}$$

requires all possible values of x_1+x_2 to always equal 3. As opposed to this, a goal requiring $x_1+x_2=3$ is not mandatory, and we can choose values of $x_1+x_2 \geq 3$ as well as $x_1+x_2 \leq 3$. In a goal constraint, positive and negative deviational variables are introduced to represent constraint violations as follows:

$$x_1+ x_2+d_1^- - d_1^+ = 3 \qquad d_1^+, d_1^- \geq 0 \tag{4.20}$$

Note that, if $d_1^- > 0$, then $x_1+x_2<3$, and if $d_1^+ > 0$, then $x_1+x_2>3$

By assigning suitable weights w_1^- and w_1^+ on d_1^- and d_1^+ in the objective function, the model will try to achieve the sum x_1+x_2 as close as possible to 3. If the goal were to satisfy $x_1+x_2\geq 3$, then only d_1^- is assigned a positive weight in the objective, while the weight on d_1^+ is set to zero.

4.4.3.1 General GP Model

A general multiple criteria mathematical programming (MCMP) problem is given as follow:

$$\begin{aligned} &\text{Max} \quad \mathbf{F}(\mathbf{x}) = \{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})\} \\ &\text{Subject to} \quad g_j(\mathbf{x}) \leq 0 \text{ for } j=1, \ldots, m \end{aligned} \tag{4.21}$$

where $\mathbf{x}$ is an n-vector of *decision variables* and $f_i(\mathbf{x})$, $i=1, \ldots, k$ are the *k criteria/objective functions*.

$$\text{Let } S=\{\mathbf{x}/g_j(\mathbf{x}) \leq 0, \text{ for all "}j\text{"}\}$$

$$Y=\{\mathbf{y}/\mathbf{F}(\mathbf{x})=\mathbf{y} \text{ for some } \mathbf{x}\in S\}$$

S is called the decision space and *Y* is called the criteria or objective space in MCMP.

Consider the general MCMP problem given in Equation 4.21. The assumption that there exists an optimal solution to the MCMP problem involving multiple criteria implies the existence of some preference ordering of the criteria by the decision maker (DM). The GP formulation of the MCMP problem requires the DM to specify an acceptable level of achievement (b_i) for each criterion f_i and specify a weight w_i (ordinal or cardinal) to be associated with the deviation between f_i and b_i. Thus, the GP model of an MCMP problem becomes:

$$\text{Minimize } Z = \sum_{i=1}^{k} (w_i^+ \ d_i^+ + w_i^- \ d_i^-) \tag{4.22}$$

$$\text{Subject to: } f_i \ (\text{x}) \ + \ d_i^- - d_i^+ = b_i \text{ for } i = 1, ..., k \tag{4.23}$$

$$g_j(x) \leq 0 \quad \text{for } j = 1, \ldots, m \tag{4.24}$$

$$x_j, \ d_i, \ d_i^+ \ \geq 0 \text{ for all } i \text{ and } j \tag{4.25}$$

Equation 4.22 represents the objective function of the GP model, which minimizes the weighted sum of the deviational variables. The system of equations (Equation 4.23) represents the goal constraints relating the multiple criteria to the goals/targets for those criteria. The variables, d_i^- and d_i^+, in Equation 4.23 are the deviational variables, representing the under achievement and over achievement of the i^{th} goal. The set of weights (w_i^+ and w_i^-) may take two forms:

1. Prespecified weights (cardinal).
2. Preemptive priorities (ordinal).

Under prespecified (cardinal) weights, specific values in a relative scale are assigned to w_i^+ and w_i^- representing the DM's "trade-off" among the goals. Once w_i^+ and w_i^- are specified, the goal program represented by Equations 4.22 through 4.25 reduces to a single objective optimization problem. The cardinal weights could be obtained from the DM using any of the methods discussed in Section 4.3 (rating, Borda count and AHP). However, for this method to work effectively, criteria values have to be scaled properly. In reality, goals are usually incompatible (i.e. in commensurable) and some goals can be achieved only at the expense of some other goals. Hence, preemptive GP, which is more common in practice, uses ordinal ranking or preemptive priorities to the goals by assigning incommensurable goals to different priority levels and weights to goals at the same priority level. In this case, the objective function of the GP model (Equation 4.22) takes the form

$$\text{Minimize } Z = \sum_{p} P_p \sum_{i} (w_{ip}^+ \ d_i^+ + w_{ip}^- \ d_i^-) \tag{4.26}$$

where P_p represents priority p with the assumption that P_p is much larger then P_{p+1} and w_{ip}^+ and w_{ip}^- are the weights assigned to the i^{th} deviational variables at priority p. In this manner, lower priority goals are considered only after attaining the higher priority goals. Thus, preemptive GP is essentially a sequential single objective optimization process, in which successive optimizations are carried out on the alternate optimal solutions of the previously optimized goals at higher priority. In addition to preemptive and nonpreemptive GP models, other approaches (fuzzy GP, min–max GP) have also been proposed. In the next four sections we will illustrate four different variants of GP for the supplier selection model discussed in Section 4.4.2.

4.4.4 Preemptive GP

For the three criteria supplier order allocation problem (Section 4.4.2), the preemptive GP formulation will be as follows:

$$\min P_1 d_1^+ + P_2 d_2^+ + P_3 d_3^+ \tag{4.27}$$

Subject to:

$$\sum_i \sum_j \sum_k \sum_m l_{ijk} \cdot x_{ijkm} + d_1^- - d_1^+ = \text{Lead time goal} \tag{4.28}$$

$$\left(\sum_i \sum_j \sum_k \sum_m p_{ikm} \cdot x_{ijkm} + \sum_k F_k . z_k \right) + d_2^- - d_2^+ = \text{Price goal} \tag{4.29}$$

$$\sum_i \sum_j \sum_k \sum_m q_{ik} \cdot x_{ijkm} + d_3^- - d_3^+ = \text{Quality goal} \tag{4.30}$$

$$d_n^-, d_n^+ \geq 0 \quad \forall n \in \{1, \ldots, 3\} \tag{4.31}$$

$$\sum_j \sum_m x_{ijkm} \leq \text{CAP}_{ik} \cdot z_k \quad \forall i, k \tag{4.32}$$

$$\sum_k \sum_m x_{ijkm} = D_{ij} \quad \forall i, j \tag{4.33}$$

$$\sum_k z_k \leq N \tag{4.34}$$

$$x_{ijkm} \leq (b_{ikm} - b_{ik(m-1)}) \cdot y_{ijkm} \quad \forall i, j, k \quad 1 \leq m \leq m_k \tag{4.35}$$

$$x_{ijkm} \geq (b_{ikm} - b_{ik(m-1)}) \cdot y_{ijkm(m+1)} \quad \forall i, j, k \quad 1 \leq m \leq m_k - 1 \tag{4.36}$$

$$x_{ijkm} \geq 0 \quad z_k \in \{0,1\} \quad y_{ijkm} \in \{0,1\} \tag{4.37}$$

4.4.5 Nonpreemptive GP

In nonpreemptive GP model, the buyer sets goals to achieve for each objective and preferences in achieving those goals expressed as numerical weights. In the nonpreemptive GP the buyer has three goals, namely,

- Limit the lead-time to lead goal with weight w_1.
- Limit the total purchasing cost to price goal with weight w_2.
- Limit the quality to quality goal with weight w_3.

The weights w_1, w_2 and w_3 can be obtained using the methods discussed in Section 4.3. The nonpreemptive GP model can be formulated as:

$$\text{Min } Z = w_1 * d_1^+ + w_2 * d_2^+ + w_3 * d_3^+ \tag{4.38}$$

Subject to the constraints (Equations 4.28 through 4.37).

In the above model, d_1^+, d_2^+ and d_3^+ represent the overachievement of the stated goals. Due to the use of the weights the model needs to be scaled. The weights w_1, w_2 and w_3 can be varied to obtain different GP optimal solutions.

4.4.6 Tchebycheff (min–max) GP

In this GP model, the DM only specifies the goals/targets for each objective. The model minimizes the maximum deviation from the stated goals. For the supplier selection problem the Tchebycheff goal program becomes:

$$\text{Min–max } (d_1^+ , d_2^+, d_3^+) \tag{4.39}$$

$$d_i^+ \geq 0 \, \forall i \tag{4.40}$$

Subject to the constraints (Equations 4.28 through 4.37).

Equation 4.39 can be reformulated as a linear objective by setting

$$\text{Max } (d_1^+, d_2^+, d_3^+) = M \geq 0$$

Thus Equation 4.39 is equivalent to:

$$\text{Min } Z = M \tag{4.41}$$

Subject to:

$$M \geq \left(d_1^+\right) \tag{4.42}$$

$$M \geq \left(d_2^+\right) \tag{4.43}$$

$$M \geq \left(d_3^+\right) \tag{4.44}$$

$$d_i^+ \geq 0 \, \forall i \tag{4.45}$$

Constraints 4.28 through 4.37 stated earlier will also be included in this model.

The advantage of Tchebycheff goal program is that there is no need to get preference information (priorities or weights) about goal achievements from the DM. Moreover, the problem reduces to a single objective optimization problem. The disadvantages of this method are (i) the scaling of goals is necessary (as required in nonpreemptive GP) and (ii) outliers are given more importance and could produce poor solutions.

4.4.7 Fuzzy GP

Fuzzy GP uses the ideal values as targets and minimizes the maximum normalized distance from the ideal solution for each objective. An ideal solution is the vector of best values of each criterion obtained by optimizing each criterion independently ignoring other criteria. In this example, ideal solution is obtained by minimizing price, lead-time and quality independently. In most situations ideal solution is an infeasible solution since the criteria conflict with one another.

If M equals the maximum deviation from the ideal solution, then the fuzzy GP model is as follows:

$$\text{Min } Z = M \tag{4.46}$$

Subject to:

$$M \geq \left(d_1^+\right)/\lambda_1 \tag{4.47}$$

$$M \geq \left(d_2^+\right)/\lambda_2 \tag{4.48}$$

$$M \geq \left(d_3^+\right)/\lambda_3 \tag{4.49}$$

$$d_i^+ \geq 0 \; \forall i \tag{4.50}$$

Constraints 4.28 through 4.37 stated earlier will also be included in this model, except that the target for Equations 4.28 through 4.30 are set to their respective ideal values. In the above model λ_1, λ_2 and λ_3 are scaling constants to be set by the user. A common practice is to set the values $\lambda_1, \lambda_2, \lambda_3$ equal to the respective ideal values. The advantage of fuzzy GP is that no target values have to be specified by the DM.

For additional readings on the variants of fuzzy GP models, the reader is referred to Ignizio and Cavalier [75], Tiwari et al. [76,77], Mohammed [78], and Hu et al. [79].

An excellent source of reference for GP methods and applications is the textbook by Schniederjans [80].

We shall now illustrate the four GP methods using a supplier order allocation case study. The data used in all four methods is presented next,

4.4.8 Case Study

To demonstrate the use of GP in supplier selection, consider the case where we have two products, one buyer, five suppliers where each supplier offers two price breaks. The problem here is to find which supplier(s) to buy from and how much to buy from the chosen supplier(s). The GP problems are solved using LINGO 8.0.

The cost of acquiring one unit of demand for product i from supplier k at price level m, p_{ikm}, is given in Table 4.16. The quantity at which at which price break occurs for product i for supplier k, b_{ikm}, is given in Table 4.17. Fixed cost associated with supplier k, F_k, is given in Table 4.18. The demand of product i by buyer j, d_{ij}, is given in Table 4.19. Lead time for supplier k to produce and supply product i to buyer j, l_{ijk}, is given in Table 4.20. Quality that supplier k maintains for product i, q_{ik}, is given in Table 4.21. Production capacity for supplier k for product i, CAP_{ik}, is given in Table 4.22.

The maximum number of suppliers that can be selected is assumed as 3.

All four GP models (preemptive, nonpreemptive, Tchebycheff and fuzzy) are used to solve the supplier order allocation problem. Each model produces a different optimal solution. They are discussed next.

TABLE 4.16 Unit Price for Supplier Product Combination

Product	Supplier	Break	Unit Price
1	1	1	190
1	1	2	175
1	2	1	200
1	2	2	170
1	3	1	185
1	3	2	177
1	4	1	188
1	4	2	180
1	5	1	194
1	5	2	172
2	1	1	360
2	1	2	335
2	2	1	370
2	2	2	330
2	3	1	355
2	3	2	340
2	4	1	365
2	4	2	337
2	5	1	357
2	5	2	350

TABLE 4.17 Price Break Data

Product	Supplier	Break	Quantity
1	1	1	90
1	1	2	200
1	2	1	80
1	2	2	180
1	3	1	100
1	3	2	180
1	4	1	85
1	4	2	170
1	5	1	90
1	5	2	168
2	1	1	200
2	1	2	350
2	2	1	210
2	2	2	330
2	3	1	220
2	3	2	338
2	4	1	180
2	4	2	400
2	5	1	177
2	5	2	365

4.4.8.1 Preemptive GP Solution

In preemptive GP, lead-time is given the highest priority, followed by price and quality, respectively. The target values for each of the objectives are set at 105% of the ideal value. For example, the ideal (minimum) value for price objective is $201,590; hence the target value for price is $211,669 and the goal is to

TABLE 4.18 Fixed Supplier Cost

Supplier	Cost
1	1000
2	1500
3	800
4	1600
5	1100

TABLE 4.19 Demand Data

Product	Buyer	Demand
1	1	320
2	1	230

TABLE 4.20 Lead-time Data

Product	Buyer	Supplier	Lead-time
1	1	1	6
1	1	2	10
1	1	3	7
1	1	4	14
1	1	5	5
2	1	1	11
2	1	2	6
2	1	3	7
2	1	4	6
2	1	5	9

TABLE 4.21 Supplier Quality Data

Product	Supplier	Quality
1	1	0.03
1	2	0.04
1	3	0.08
1	4	0.09
1	5	0.06
2	1	0.06
2	2	0.08
2	3	0.04
2	4	0.03
2	5	0.03

minimize the deviation above the target value. Table 4.23 illustrates the solution using the preemptive GP model.

4.4.8.2 Nonpreemptive GP

In nonpreemptive GP, weights w_1, w_2 and w_3 are obtained using AHP. The values of the weights for lead-time, price and quality are assumed to be 0.637, 0.185 and 0.178, respectively. The target values

TABLE 4.22 Production Capacity Data

Product	Supplier	Capacity
1	1	450
1	2	400
1	3	470
1	4	350
1	5	500
2	1	600
2	2	550
2	3	480
2	4	590
2	5	640

TABLE 4.23 Preemptive GP Solution

Preemptive GP	Ideal Values	Preemptive Priorities	Target for Preemptive Goal (Ideal + 5%)	Actual Achieved	Whether Goal Achieved	Suppliers Chosen
Lead-time	4610	1	4840	4648	Achieved	S1,S4,S5
Price	201590	2	211669	203302	Achieved	
Quality	25.2	3	26.46	30.59	Not achieved	

TABLE 4.24 Nonpreemptive GP Solution

Nonpreemptive GP	Ideal Values	Weights	Scaling Sonstant	Target for Nonpreemptive Goal (Ideal + 5%)	Actual Achieved	Whether Goal Achieved	Suppliers Chosen
Lead-time	4610	0.637	4840	4840	4840	Achieved	S1,S2,S3
Price	201590	0.185	211669	211669	204582	Achieved	
Quality	25.2	0.178	26.46	26.46	29.76	Not achieved	

TABLE 4.25 Tchebycheff GP Solution

Tchebycheff GP	Ideal Values	Weights	Scaling Constant	Targets for Tchebycheff GP	Actual Achieved	Whether Goal Achieved	Suppliers Chosen
Lead-time	4610	0.637	0.1	4840	4932	Not achieved	S1,S4,S5
Price	201590	0.185	0.001	211669	205196	Achieved	
Quality	25.2	0.178	10	26.46	29.76	Not achieved	

used are the same that are used in preemptive GP. The solution of the nonpreemptive model is shown in Table 4.24. Since nonpreemptive GP requires scaling, the target values are used as scaling constants.

4.4.8.3 Tchebycheff GP

In Tchebycheff GP, the target values are the same that are used in preemptive GP. Scaling constants are chosen in such a way that all the three objectives have a similar magnitude. For example lead-time when multiplied by 0.1 gives 461 and quality when multiplied by 10 yields 252; this makes lead-time and quality of similar magnitude. Using the Tchebycheff method, the solution obtained is illustrated in Table 4.25.

TABLE 4.26 Fuzzy GP Solution

Fuzzy GP	Ideal Values	Weights	Scaling Constant	Target for Fuzzy GP	Actual Achieved	Whether Goal Achieved	Suppliers Chosen
Lead-time	4610	0.637	0.1	4610	4675	Not achieved	S1,S2,S5
Price	201590	0.185	0.001	201590	206056	Not achieved	
Quality	25.2	0.178	10	25.2	27.53	Not achieved	

4.4.8.4 Fuzzy GP

Recall that in fuzzy GP, the ideal values are used as targets for the different goals. The solution obtained using fuzzy GP is shown in Table 4.26. The scaling constants are calculated such that all the objectives have similar values.

Here the final set of suppliers is different since the targets are set at ideal values and it is generally not possible to achieve any of them.

4.4.9 Value Path Approach

The presentation of results presents a critical link, in any multiobjective problem. Any sophisticated analysis, just becomes numbers, if they are not presented to the decision maker in an effective way. In case of multiobjective problems, a lot of information needs to be conveyed which not only includes performance of various criteria but also their trade-offs. In this section, we discuss how to present the four different optimal solutions obtained by the different GP models.

The value path approach [81] is one of the most efficient ways to demonstrate the trade-offs among the criteria obtained by the different solutions. The display consists of a set of parallel scales; one for each criterion, on which is drawn the value path for each of the solution alternative. Value paths have proven to be an effective way to present the trade-offs in problems with more than two objectives. The value assigned to each solution on a particular axis is that solutions value for the appropriate objective divided by the best solution for that objective. Therefore the minimum value for each axis is 1. Following are some properties of the value path approach [81]:

- If two value paths representing solutions A and B intersect between two vertical scales, then the line segment connecting A and B in objective space has a negative slope and neither objective dominates other.
- If three or more value paths intersect, then their associated points in the objective space are collinear.
- If two paths do not intersect, then one path must lie entirely below the other and is therefore inferior.
- Given, any intersecting pair of value paths, a third value path is inferior if it does not lie above the intersecting point.

4.4.9.1 Value Path Approach for the Supplier Selection Case Study

The supplier selection case study was solved using four different GP approaches as previously illustrated. A summary of the results is provided in Table 4.27.

To present these results to the decision maker, value path approach is used as follows.

1. Find the best (minimum) value obtained for each criterion. For the price criterion, preemptive GP has the best value of 203302; for the lead-time criterion Preemptive GP is best with a value of 4648; and for the quality objective the best value of 27.53 is obtained through fuzzy GP.
2. For each solution, divide their objective values by the best value for that objective. For example, for the *preemptive GP* solution, the value of lead-time, price and quality are 4648, 203302 and

TABLE 4.27 Summary of GP Solutions for the Case Study

Solution		Lead-time	Price	Quality
1	Preemptive GP	4648	203302	30.59
2	Nonpreemptive GP	4840	204582	29.76
3	Tchebycheff GP	4932	205196	29.76
4	Fuzzy GP	4675	206056	27.53

TABLE 4.28 Criteria Values for Value Path Approach

Alternative	Method	Lead-time	Price	Quality
1	Preemptive GP	1	1	1.1112
2	Nonpreemptive GP	1.0413081	1.0063	1.081
3	Tchebycheff GP	1.0611015	1.00932	1.081
4	Fuzzy GP	1.005809	1.01355	1

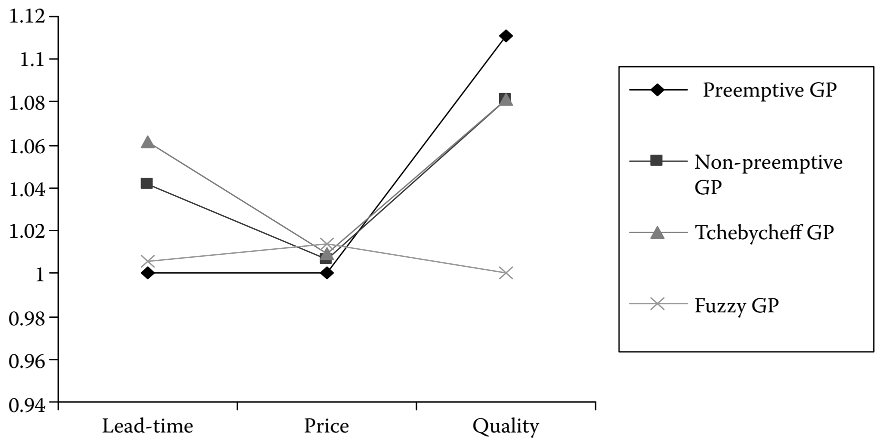


FIGURE 4.7 Graphical representation of value path approach.

30.59, respectively. The best values for lead-time, price and quality are 4648, 203302 and 27.53. Therefore the values for the value path approach corresponding to *preemptive GP* are obtained as (4648/4648), (203302/203302) and (30.59/27.53), respectively. Similar values are calculated for the other solutions under the value path approach as shown in Table 4.28.

3. Plot the results, with lead-time, price and quality on X-axis and ratios of the objective values on the Y-axis.

The last step is to plot them on a graph. The graph is shown in Figure 4.7.

4.4.9.2 Discussion of Value Path Results

Based upon the preference of the decision maker, the preferred suppliers and the quantity ordered from each can change. The value path approach is a useful tool to compare the trade-offs among the suppliers. In some cases, the price of the product may dictate the suppliers who are chosen and in some other cases, the suppliers chosen may be dictated by the lead-time or the quality. Hence, value path approach can be used to study the trade-offs between different solutions. For example, from Figure 4.7 it can be seen that preemptive GP does 1% better on lead-time and 2% better on price compared to fuzzy GP, but fuzzy GP is 11% better on quality; such comparisons can easily demonstrate the trade-offs between different solutions.

4.5 Concluding Remarks

While supplier selection plays an important role in purchasing, it is especially important for the DOD which spends billions of dollars on procurement. This chapter illustrates the use of both discrete and continuous MCDM techniques to optimize the supplier selection process. In this chapter, we present the supplier selection problem in two phases. In the first phase called, prequalification, we reduce the initial set of large number suppliers to a manageable set. Phase one reduces the effort of the buyer and makes the prequalification process entirely objective. In the second phase, we analyze the shortlisted suppliers using the multiobjective technique known as goal programming. We consider several conflicting criteria, including, price, lead-time and quality. An important distinction of multiobjective techniques is that it does not provide one optimal solution, but a number of solutions known as efficient solutions. Hence, the role of the decision maker (buyer) is more important than before. By involving the decision maker early in the process, the acceptance of the model results by the top management becomes easier. The efficient solutions are compared using the value path approach to show the criteria trade off obtained using different GP approaches. Besides GP, there are other approaches to solve the multicriteria optimization model for the supplier selection problem formulated in Section 4.4.2. They include weighted objective method, compromise programming and interactive approaches. Interested readers can refer to Wadhwa and Ravindran [1] and Ravindran [71] for more details. Reference [71] also provides information on the computer software available for MCDM methods.

The supplier selection models can be extended in many different directions. One of the areas is managing supplier risk. Along with cost, quality, technology and service criteria used in sourcing decisions, there is a need to integrate global risk factors such as political stability, currency fluctuations, taxes, local content rules, infrastructure (ports, raw materials, communication, transportation, certification, etc.) in the supplier selection process. Another area of research is *supplier monitoring*. Since the supplier performance factors can change over time, real time monitoring of suppliers becomes critical. The monitoring issues are to determine what supplier performance factors to monitor and the frequency of monitoring.

References

1. Wadhwa, V. and Ravindran, A. 2007. Vendor selection in outsourcing. *Computers & Operations Research*, 34, 3725–3737.
2. Aberdeen Group. 2004. Outsourcing portions of procurement now a core strategy. *Supplier Selection and Management Report*, 4(7), 4.
3. Sonmez, M. 2006. A review and critique of supplier selection process and practices, 3–34. Loughborough University Business School, UK.
4. Weber, C.A., Current J.R., and Benton, W.C. 1991. Vendor selection criteria and methods. *European Journal of Operational Research*, 50, 2–18.
5. Stanley, E.D., Honig, D.P., and Gainen, L. 1954. Linear programming in bid evaluation. *Naval Research Logistics Quarterly*, 1, 365–368.
6. National Bureau of Standards. 1954. SEAC determines low bidders. *Technical News Bulletin*, 38–179.
7. Gainen, L. 1955. Linear programming in bid evaluation. *Proceedings of the Second Symposium of linear Programming DC/S Comptroller*, USAF, Washington, DC, 1, 19–38.
8. Deng, S.J. and Elmaghraby, W. 2005. Supplier selection via tournaments. *Production and Operations Management*, 14, 252–267.
9. Drewes, R. 1987. *The Air Force and the Great EngineWar*. National Defense University Press, Washington, DC.
10. Office of Management and Budget, Department of Defense. Available at http://www.whitehouse.gov/omb/budget/fy2008/defense.html. Accessed 1/19/08.

11. Aissaoui, N., Haouari, M., and Hassini, E. 2007. Supplier selection and order lot sizing modeling: A review. *Computers & Operations Research*, 34(12), 3516–3540.
12. De Boer, L., Labro, E., and Morlacchi, P. 2001. A review of methods supporting supplier selection. *European Journal of Purchasing and Supply Management,* 7, 75–89.
13. Dickson, G. W. 1966. An analysis of vendor selection systems and decisions. *Journal of Purchasing*, 2, 5.
14. Wilson, E.J. 1994. The relative importance of supplier selection. *International Journal of Purchasing and Materials Management*, 30, 35–41.
15. Mandal, A. and Deshmukh, S.G. 1994. Vendor selection using interpretive structural modeling (ISM). *International Journal of Operations and Production Management*, 14(6), 52–59.
16. Vokurka, R.J., Choobineh, J, and Vadi, L. 1996. A prototype expert system for the evaluation and selection of potential suppliers. *International Journal of Operations and Production Management*, 16(12), 106–127.
17. Goicoechea, A., Hansen, D.R., and Duckstein, L. 1982. *Multiobjective Decision Analysis with Engineering and Business Applications*. Wiley, New York city, NY.
18. Weber C.A. and Ellram, L.M. 1992. Supplier selection using multi-objective programming a decision support system approach. *International Journal of Physical Distribution and Logistics Management*, 23(2), 3–14.
19. Weber, C.A. Current, J.R. and Desai, A. 2000. An optimization approach to determining the number of suppliers to employ. *Supply Chain Management: An International Journal*, 2(5), 90–98.
20. Hinkel, C.L., Robinson, P.J., and Green, P.E. 1969. Vendor evaluation using cluster analysis. *Journal of Purchasing*, 5(3), 49–58.
21. Holt, G.D. 1998. Which contractor selection methodology? *International Journal of Project Management*, 16(3), 153–164.
22. Ng, S.T. and Skitmore, R.M. 1995. CP-DSS: decision support system for contractor prequalification. *Civil Engineering Systems: Decision Making Problem Solving*, 12(2), 133–160.
23. Mendoza, A., Ravindran, A., and Santiago, E. 2008. A three phase multi-criteria method to the supplier selection problem. *International Journal of Industrial Engineering*, 15(2). (In press)
24. Ghodsypour, S.H. and O'Brien, C. 2001.The total cost of logistics in supplier selection, under conditions of multiple sourcing, multiple criteria and capacity constraint. *International Journal of Production Economics,* 73(1), 15–28.
25. Wind, Y. and Robinson, P.J. 1968. The determinants of vendor selection: evaluation function approach, *Journal of Purchasing and Materials Management,* 4, 29–41.
26. Zenz, G. 1981. Purchasing and the Management of Materials. Wiley, New York.
27. Timmerman, E. 1986. An approach to vendor performance evaluation. *Journal of Purchasing and Supply Management*, 1, 2–8.
28. Saaty, T.L. 1980. *The Analytic Hierarchy Process.* McGraw Hill, New York city, NY.
29. Akarte, M. M., Surendra, N. V., Ravi, B., and Rangaraj, N. 2001. Web based casting supplier evaluation using analytical hierarchy process. *Journal of the Operational Research Society*, 52, 511–522.
30. Barbarosoglu, G. and Yazgac, T. 1997. An application of the analytic hierarchy process to the supplier selection problem. *Product and Inventory Management Journal*, 38(1), 14–21.
31. Bharadwaj, N. 2004. Investigating the decision criteria used in electronic components procurement. *Industrial Marketing Management*, 33(4). 317–323.
32. Ghodyspour, S.H. and O'Brien, C. 1998. A decision support system for supplier selection using an integrated analytic hierarchy process and linear programming. *International Journal of Production Economics*, 56–57, 199–212.
33. Muralidharan, C., Anantharaman, N., and Deshmukh, S.G. 2001. Vendor rating in purchasing scenario: a confidence interval approach. *International Journal of Operations and Production Management*, 21(10), 22–33.
34. Nydick, R.L. and Hill, R.P. 1992. Using the Analytic Hierarchy Process to structure the supplier selection procedure. *International Journal of Purchasing and Materials Management*, 28(2), 31–36.

35. Smytka, D.L. and Clemens, M.W. 1993. Total cost supplier selection model: A case study. *International Journal of Purchasing and Materials Management*, 29(1), 42–49.
36. Ellram, L.M. 1995. Total cost of ownership, an analysis approach for purchasing. *International Journal of Physical Distribution and Logistics Management,* 25, 4–23.
37. Degraeve, Z. and Roodhooft, F. 1999. Improving the efficiency of the purchasing process using total cost of ownership information: The case of heating electrodes at Cockerill Sambre S.A. *European Journal of Operational Research*, 112, 42–53.
38. Degraeve, Z. and Roodhooft, F. 2000. A mathematical programming approach for procurement using activity based costing. *Journal of Business Finance and Accounting,* 27, 69–98.
39. Degraeve, Z, Labro, E., and Roodhooft, F. 2004 Total cost of ownership purchasing of a service: the case of airline selection at Alcatel Bell. *European Journal of Operational Research*, 156, 23–40.
40. Moore, D.L. and Fearon, H.E. 1973. Computer-assisted decision-making in purchasing, *Journal of Purchasing*, 9, 5–25.
41. Gaballa, A.A. 1974. Minimum cost allocation of tenders. *Operational Research Quarterly*, 25, 389–398.
42. Anthony, T.F. and Buffa, F.P. 1977. Strategic purchasing scheduling. *Journal of Purchasing and Materials Management*, 13, 27–31.
43. Bender, P.S., Brown, R.W., Isaac, M. H., and Shapiro, J.F. 1985. Improving purchasing productivity at IBM with a normative decision support system. *Interfaces*, 15, 106–115.
44. Narasimhan, R. and Stoynoff, K. 1986. Optimizing aggregate procurement allocation decisions. *Journal of Purchasing and Materials Management*, 22, 23–30.
45. Turner, I. 1988. An independent system for the evaluation of contract tenders. *Operational Research Society*, 39, 551–561.
46. Pan, A.C. 1989. Allocation of order quantity among suppliers. *Journal of Purchasing and Materials Management*, 25, 36–39.
47. Sheshadri, S., Chatterrjee, K., and Lilien, G.L. 1991. Multiple source procurement competitions. *Marketing Science*, 10(3), 246–253.
48. Benton, W.C. 1991. Quantity discount decision under conditions of multiple items. *International Journal of Production Research*, 29, 1953–1961.
49. Chaudhry, S.S., Forst, F.G., and Zydiak, J.L. 1991. Multicriteria approach to allocating order quantities among suppliers. *Production and Inventory Management Journal*; 3rd Quarter, 82–86.
50. Jayaraman, V., Srivastava, R., and Benton, W.C. 1999. Supplier selection and order quantity allocation: A comprehensive model. *Journal of Supply Chain Management*, 35(2), 50–58.
51. Feng, C., Wang, X., and Wang, J.S. 2001. An optimization model for concurrent selection of tolerances and suppliers. *Computers and Industrial Engineering*, 40, 15–33.
52. Buffa, F.P. and Jackson, W.M. 1983. A goal programming model for purchase planning. *Journal of Purchasing and Materials Management*, Fall, 27–34.
53. Sharma, D., Benton, W.C., and Srivastava, R. 1989. Competitive strategy and purchasing decisions. *Proceedings of the 1989 Annual Conference of the Decision Sciences Institute*, 1088–1090.
54. Liu F, Ding F.Y., and Lall, V. 2000. Using data envelopment analysis to compare suppliers for supplier selection and performance improvement. *Supply Chain Management: An International Journal*, 5(3), 143–150.
55. Karpak B., Kumcu, E., and Kasuganti, R. 1999. An application of visual interactive goal programming: A case in supplier selection decisions. *Journal of Multi-criterion Decision Making*, 8, 93–105.
56. Bhutta, K.S. and Huq, F. 2002. Supplier selection problem: A comparison of the total cost of ownership and Analytical Hierarchy Process. *Supply Chain Management,* 7(3–4), 126–135.
57. Yang, T. 2006. Multiobjective optimization models for managing supply risks in supply chains. PhD Thesis, Department of Industrial Engineering, Pennsylvania State University.
58. Xia, W. and Wu, Z. 2007. Supplier selection with multiple criteria in volume discount environments. *Omega*, 35, 494–504.

59. Saen, R.F. 2007. Suppliers selection in the presence of both cardinal and ordinal data. *European Journal of Operational Research,* 183, 741–747.
60. Chen, C.-T., Lin, C.-T., and Huang, S.-F. 2006. A fuzzy approach for supplier evaluation and selection in supply chain management. *International Journal of Production Economics*, 102, 289–301.
61. Haq, A.N. and Kannan, G. 2006. Design of an integrated supplier selection and multi-echelon distribution inventory model in a built-to-order supply chain environment. *International Journal of Production Research*, 44, 1963–1985.
62. Cachon, G.P. and Zhang, F. 2006. Procuring fast delivery: Sole sourcing with information asymmetry. *Management Science*, 52, 881–896.
63. Ding, H., Benyoucef, L., and Xie, X. 2005. A simulation optimization methodology for supplier selection problem. *International Journal of Computer Integrated Manufacturing*, 18, 210–224.
64. Piramuthu, S. 2005. Knowledge-based framework for automated dynamic supply chain configuration. *European Journal of Operational Research*, 165, 219–230.
65. Liu, J. and Wu, C. 2005. "An integrated method for supplier selection in SCM," *In Proceedings of the* ICSSSM 35, Vol. 1, pp. 617–620. Chongging, China: Publisher IEEE.
66. Valluri, A. and Croson, D. 2005. Agent learning in supplier selection models. *Decision Support Systems*, 39, 219–240.
67. Emerson, D. and Piramuthu, S. 2004. Agent-based framework for dynamic supply chain configuration, In *Proceedings of the 37th Hawaii International Conference on System Sciences,* pp. 1–8. Big Island, HI, USA: Publisher IEEE.
68. Agrell, P.J., Lindroth, R., and Norman, A. 2004. Risk, information and incentives in telecom supply chains. *International Journal of Production Economics*, 90, 1–16.
69. Chan, F.T.S. 2003. Interactive selection model for supplier selection process: An analytical hierarchy process approach. *International Journal of Production Research*, 41, 3549–3579.
70. Choy, K.L., Lee, W.B., and Lo, V. 2003. Design of an intelligent supplier relationship management system: A hybrid case based neural network approach. *Expert Systems with Applications*, 24, 225–237.
71. Ravindran, A.R. 2008. *Operations Research and Management Science Handbook*. CRC Press, Boca Raton, FL.
72. Khattree, R., and Dayanand, N. 2000. *Multivariate Data Reduction and Discrimination with SAS Software*. SAS Institute Inc., Cary, NC.
73. Powdrell, B.J. 2003. Comparison of MCDM algorithms for discrete alternatives. Masterís Thesis, Department of Industrial Engineering, Pennsylvania State University.
74. Patel, U.R. 2007. Experiments in group decision making in the Analytic Hierarchy Process. Masterís Thesis, Department of Industrial Engineering, Pennsylvania State University.
75. Ignizio, J.M., and Cavalier, T.M. 1994. *Linear Programming*. Prentice Hall, Englewood Cliffs, NJ.
76. Tiwari, R.N., Dharmar, S., and Rao, J.R. 1986. Priority structure in fuzzy goal programming. *Fuzzy Sets and Systems*, 19, 251–259.
77. Tiwari, R.N., Dharmar, S., and Rao, J.R. 1987. Fuzzy goal programming—an additive model. *Fuzzy Sets and Systems*, 24, 27–34.
78. Mohammed, R.H. 1997. The relationship between goal programming and fuzzy programming. *Fuzzy Sets and Systems*, 89, 215–222.
79. Hu, C.F., Teng, C.J., and Li, S.Y. 2007. A fuzzy goal programming approach to multiobjective optimization problem with priorities. *European Journal of Operational Research*, 176, 1319–1333.
80. Schniederjans, M. 1995. *Goal Programming: Methodology and Applications.* Kluwer Academic Publishers, Boston, MA.
81. Schilling, D.A., Revelle, C., and Cohon, J. 1983. An approach to the display and analysis of multiobjective problems. *Socio-economic Planning Sciences,* 17(2), 57–63.

5

Probabilistic Modeling for UAV Path Planning in the Presence of Threat Zones

Barbara Pfeiffer
University of Wuppertal

Rajan Batta
University at Buffalo

Kathrin Klamroth
University of Wuppertal

Rakesh Nagi
University at Buffalo

5.1 Introduction

Unmanned aerial vehicles (UAVs) are used for military as well as commercial interests. In a military context the role of UAVs is, among others, to provide the operator with valuable intelligence about an area of interest, to gather reconnaissance information and surveillance data, and during execution to offer support in operations against enemy forces or integrated air defense systems. UAVs have already demonstrated that they possess outstanding capability on these fields. According to Nielsen et al. (2003) UAVs will become more and more integrated into daily operations in the next decades.

To fulfill a mission with the help of UAVs, considerable planning has to take place well in advance of the operation before any military deployment. Beside conceptual and technical aspects, this involves path planning for the UAV on the operation site which carefully incorporates the risks arising from flying through insecure and adversary regions. In order to address uncertainties arising from threat sources like radar and antiaircraft systems (see, for example, Marine Corps 2003) probabilistic path planning has become a new focus in the recent literature.

An instant of threat occurring in the range of the threat sources is regarded as an event of consequence during a mission of a UAV, i.e. a detection. The treatment of such events of consequences is a well-studied problem in the context of hazardous materials truck routing (see the recent review by Erkut and Ingolfsson 2005), where such an event corresponds to an accident of the truck. In a street network each link is associated with two attributes, the probability and the measure of the consequence of a release accident. A path is sought that minimizes a path evaluation function which is typically the expected value of the consequence of a hazardous materials truck traveling along the path (cf. Jin et al. 1996).

There are two main differences in comparison to UAV mission planning. First, a UAV facing an event of consequence may be able to continue on its route because it has just been detected, not necessarily disabled. The second difference consists in the discretization due to the network. The goal in UAV mission planning is to consider the path planning as well as the handling of the threats in a continuous setting.

For that reason different approaches have been suggested for UAV path planning. Based on GIS, Casas et al. (2005) generate a real-time network in a hostile setting adapted to the specific environment. The network is used as a reference for routing UAVs. The threat sources are considered as points in the three-dimensional space and the probability for an event of consequence on any point on a link of a network is dependent on its distance from the radars. By combining this with a nonhomogeneous Poisson process counting the number of arrivals this yields the probability of no detection along a link in the network. Optimal paths are determined with a k-shortest path algorithm where the link probabilities correspond to the weights.

Dogan (2003) utilizes a probabilistic threat exposure map, which indicates the risk of exposure to sources of threat as a function of the position. The different threat sources are described by probability density functions (PDF). An upper bound on the probability of a UAV being detected if it follows a certain path in such an area is calculated and is used as a performance index or cost function to find the minimizing path. The strategy leading a UAV to a target consists of a stepwise search starting with an initial given search direction. After each step, the PDF representing the threat zones is used to quantify the tradeoff between reducing the flight distance and threat exposure in order to determine the direction of the next step. This greedy-like strategy yields a local optimal solution.

Whereas both methods consider continuous PDFs in order to describe the threat sources the path is still discretized in the sense that Casas et al. (2005) plan on a network and Dogan and Zengin (2006) proceed in discrete steps. This is the motivation for our approach. We consider the path planning in a continuous setting from the origin to the destination. In order to obtain a model of reasonable complexity we consider a discrete approximation of the threat intensity parameter replacing the PDFs. On the resulting probabilistic map subdivided into polyhedral threat zones of constant threat, we generate a global optimal path. A completely continuous treatment of the path planning is thus still an open problem.

The remainder of the chapter is organized as follows. In Section 5.2 we review the literature related to deterministic and probabilistic approaches in path planning, and establish the motivation for our approach. Section 5.3 is devoted to the problem description and the introduction of our model for probabilistic path planning in the presence of threat. At an intuitive level, once a path for a UAV has been selected a probabilistic experiment is defined with specified values and random variables, e.g. probability of escape (a specified value), time of travel (a specified value), time till first detection (a random variable; normally we are interested in its expected value). The corresponding objectives (e.g. probability of escape, travel time, expected time till first detection) are typically conflicting. Furthermore, each candidate path yields a different probabilistic experiment with its own, usually different, objective values. In Section 5.4 we demonstrate that the probability of zero detections is an easy objective to handle, in the sense that optimization of this objective can be shown to be equivalent to a shortest path problem. Addressing a target of opportunity whose utility may change with time requires a different approach that should additionally consider the flight duration in path planning. Section 5.5 explores the flying time versus risk tradeoff using a bicriteria optimization framework. Section 5.6 discusses some realistic issues for the military context: (i) the use of the expected time until the first detection as a tie-breaker; and (ii) modeling of delay between first detection and fatal attack. Finally, we present the conclusions of this work in Section 5.7.

5.2 Literature Review

In this section we review some of the former results related to the topic of path planning for UAVs in hostile environments.

5.2.1 Deterministic Path Planning

Many papers that investigate path planning for UAVs presume that the location of the threats and their effects are deterministically known and interpret a path which avoids possible threat regions as an optimal path.

A common approach applied in most of these papers is to construct a Voronoi diagram with respect to the known locations of threats and to determine an initial solution with Dijkstraís algorithm which is then refined. Following this line of thought Bortoff (2000) proposes a two-step path-planning algorithm. In the first step, a rough-cut suboptimal path through the radar sites is constructed by generating and searching a graph based on Voronoi polygons. In the second step, the solution is improved by applying a virtual force fields approach. Therefore the path is seen as a chain of point masses and a representation of the path is generated via the steady-state equilibrium solution to this Lagrangian mechanical system driven by virtual forces. McLain et al. (2001) discretize the path by subdividing it into segments of given length and then smooth it to achieve flyable paths. Anderson et al. (2005) advance this approach and develop a real-time dynamic algorithm which generates point-constrained and time-extremal trajectories joining the discretized path segments.

5.2.2 Path Planning Involving Probabilistic Methods

Since the perception to view threat zones as no-fly areas or obstacles does not always fit real situations, there is an increasing interest in the recent literature in applying probabilistic methods to explore the uncertainties arising from threat sources which are naturally inherent in military missions in different ways.

We already discussed the strategies of Casas et al. (2005) and Dogan and Zengin (2006) that describe the threat sources with the help of PDFs. Dogan and Zengin (2006) extend their method based on the probabilistic threat exposure map also to a strategy to follow a moving target. The latter two methods can be used for real-time applications, but do not guarantee to find an optimal path.

Hespanha et al. (2001) investigate how to build probabilistic maps for threat zones based on noisy surveillance data of multiple objects. The probability map is constructed using a likelihood function which is fitted to experimental data by adapting characteristic parameters like sensor range, tracking range or errors of the sensor. It is assumed that radar sites remain at fixed locations. Based on this map, an algorithm which finds an optimal path for m given way points is presented. We note that the resulting path is a minimum risk path only among the paths which include these way points.

Jun and D'Andrea (2002) decomposes the region of interest in n cells. For each cell, occupancy rates are assumed to be known from *a priori* sensor readings and computations. Using these occupancy rates, a map assigning each cell a probability that the UAV will not be neutralized by the adversary is constructed. The problem is then to find the sequence of cells joining the origin and the destination under all possible sequences such that the probability of not being neutralized along this sequence is maximized. The problem can be equivalently formulated as a shortest path problem on a digraph, which is defined using the probabilities assigned to each cell and the knowledge about the position of adjacent cells. The vertices correspond to cells and hence a shortest path in this digraph can be identified with a sequence of cells. A polygonal path is obtained by joining the centers of the cells. This path is refined by smoothing it in order to transform it into a navigable route. Since the region is discretized, this method does not guarantee to find an optimal path.

Hallam et al. (2001) apply a multicriteria shortest path algorithm for networks in order to guide a submarine through a field of sensors within a given time period. They introduce a grid with vertices that correspond to a finite set of possible positions and edges between each pair of vertices. Related to each edge is a vector-valued cost, where one component is the time required to traverse the edge and the other component represents the (numerically evaluated) probability of being detected while traveling along this edge.

5.2.3 Critique of the Existing Models

We summarize our critique of the existing literature as follows:

- Deterministic approaches are typically threat-avoiding. They completely avoid threat zones which might neither be practical nor advantageous in all situations.
- Most of the stochastic approaches do not guarantee finding an optimal path. Often the paths are restricted by way points or based on regional discretization, or the paths are myopically developed.
- Flight time is often not considered at all. Fuel constraints and targets of opportunity cannot be handled by such approaches.
- There is seldom a distinction between UAV risk and mission objectives. In some cases, the decision maker might choose to sacrifice the UAV if mission completion is imminent.

5.3 Probabilistic Modeling of Detections Along a Path

Motivated by the above discussion, we consider the following mission planning problem. The mission of the UAV is to take out a target at a fixed destination point. In order to achieve this mission the goal is to find the best route from the given starting point of the UAV to the destination point (Figure 5.1). En route to the target the UAV passes through areas of threat. We define an instant of threat as an event of consequence, i.e. detection. For the area of the mission we assume that a probabilistic map that provides the information about the risk of exposure to instances of threat as a function of the position is available during the mission. This map is given by a finite subdivision of the continuous plane $\mathbf{R}^2$ into closed regions of constant threat (the union of all regions equals $\mathbf{R}^2$, and the intersection of the interior of each pair of such regions is empty). This kind of probabilistic map can be determined by methods described in Hespanha et al. (2001).

We assume that the UAV is goal seeking and behaves sacrificially, i.e. it is nonreactive and it stays on the chosen path whether it is threatened or not. In determining the paths to route the UAV we ignore flight dynamic considerations such as the turning radius. These assumptions are made for the sake of analytical tractability.

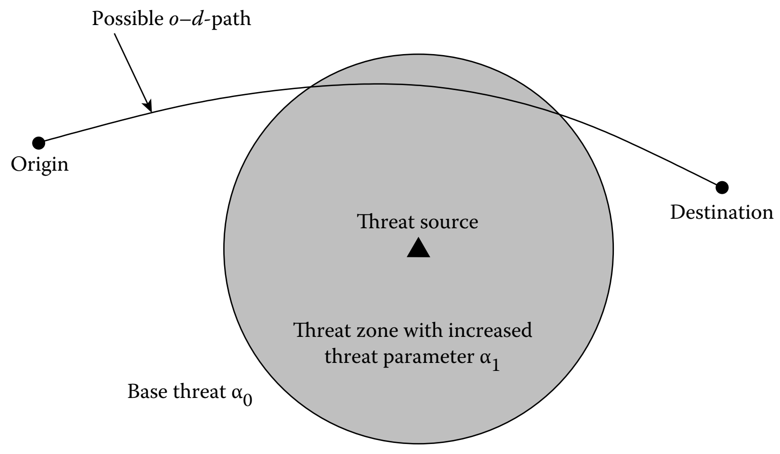


FIGURE 5.1 Possible o–d-path for UAVs in adversarial regions.

To begin our analysis we deduce a general expression for the probability of several detections along a chosen path.

Formally, given the origin point of the UAV $o \in \mathbf{R}^2$ and the destination point $d \in \mathbf{R}^2$ where the target is located, a continuous curve W given by the parameterization $w=(w_1,w_2)^T:[0,1] \rightarrow \mathbf{R}^2$ with $w(0)=o$ and $w(1)=d$ that is continuously differentiable on [0,1] with the possible exception of at most a finite number of points, where the derivative $w' = (w'_1, w'_2)^T$ has finite limits from the left and from the right, is called an o–d path. Since we consider flight routes we assume that the underlying metric is induced by the Euclidean norm $\| \cdot \|_2$. The arc length $l(W)$ of the o–d-path W is given by:

$$l(W) := \int_W \mathrm{d}q = \int_0^1 \|w'(s)\|_2 \, \mathrm{d}s.$$

Denote by $X(W)$ the random variable which counts the number of detections on an o–d-path W and by $P\{X(W)=k\}$ the probability of k detections along the path W.

Let $\lambda(x, y) \geq 0 \forall (x, y) \in \mathbf{R}^2$ be the spatially dependent threat intensity parameter which indicates the number of instants of threats as a function of the position. The average number of detections $\mu(W)$ along a path W is then given by:

$$\mu(W) := \int_W \lambda(x,y)\mathrm{d}q = \int_0^1 \lambda(w(s)) \cdot \|w'(s)\|_2 \, \mathrm{d}s. \tag{5.1}$$

Similar to models for hazardous material routing (cf. Erkut and Ingolfsson 2000) the counting process $X(W)$ for detections along a given path is a Poisson process, hence:

$$P\{X(W)=k\} = \frac{1}{k!}(\mu(W))^k e^{-\mu(W)}, k = 0, 1, 2, \ldots \tag{5.2}$$

Thus the threat parameter that indicates the rate of detection corresponds to the arrival rate in a Poisson process.

In order to build a tractable model we consider the following discrete approximation of the intensity parameters. We assume that each threat zone is the interior of a closed, bounded, pathwise connected set, i.e. for each pair of points in such a set there exists a path in the set connecting them. Each threat zone and the outside region are associated with a given threat parameter, which remains constant over the zone. The threat parameter on the boundaries of the threat zones is the minimum of the threat parameters of the adjacent zones.

We assume that g threat zones labeled TZ_m, $m=1, \ldots, g$, are given with associated threat parameters $\alpha_m \in \mathbf{R}^+ = \{x \in \mathbf{R}: x > 0\}$, $m=1, \ldots, g$. Let $\alpha_0 \in \mathbf{R}^+$ be the threat parameter outside the threat zones.

Each arbitrary o–d-path W can be decomposed into a disjunction of subpaths $W_i, i=1,\ldots,n$, $W = \biguplus_{i=1}^{n} W_i$, corresponding to the different threat zones on which the threat parameter is constant and equal to $c_i \in \{\alpha_0,\ldots,\alpha_g\}$, $i=1,\ldots,n$ (see, for example, Figure 5.2). Note that the decomposition of W into subpaths depends on the number and the structure of threat zones.

With respect to a fixed decomposition into subpaths $W = \biguplus_{i=1}^{n} W_i$ the average number of detections along W defined by Equation 5.1 is then given by:

$$\mu(W) = \int_W \lambda(x,y)\mathrm{d}q = \sum_{i=1}^{n} \int_{W_i} \lambda(x,y)\mathrm{d}q = \sum_{i=1}^{n} \int_{W_i} c_i \mathrm{d}q = \sum_{i=1}^{n} c_i \, l(W_i).$$

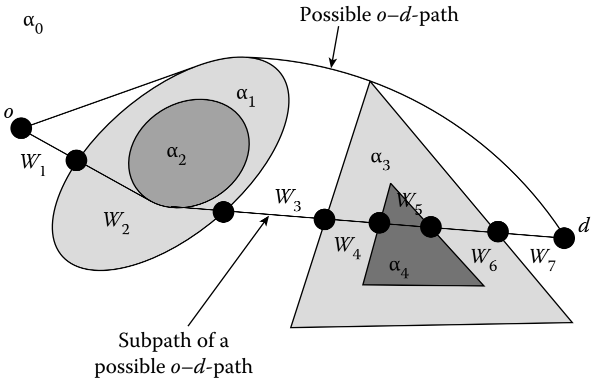


FIGURE 5.2 Illustration of threat zones.

Thus the average number of detections along W corresponds to the weighted length of the o–d-path W with respect to the threat levels c_i which is denoted by:

$$\theta := \sum_{i=1}^{n} c_i \, l(W_i).$$

Therefore the probability of less than or equal to k detections along an o–d-path W that depends on the weighted length of this path θ is given by:

$$P\{X(W) \le k\} = e^{-\theta}\left(\sum_{j=0}^{k} \frac{1}{j!}\theta^{j}\right). \tag{5.3}$$

5.4 Risk Minimization

If we assume that the value and the location of the target are time-invariant, the objective is to maximize the probability of zero events of consequence along a chosen path. The related optimization problem is:

$$\begin{aligned} &\text{maximize } P\{X(W)=0\} \\ &\text{subject to} \quad W \text{ is an } o\text{–}d\text{-path} \end{aligned} \tag{5.4}$$

Converting the maximization problem into the equivalent log-transformed minimization problem yields the correspondence to the following shortest path problem with weighted distances:

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{n} c_i l(W_i) \\ &\text{subject to} \quad W = \biguplus_{i=1}^{n} W_i. \end{aligned} \tag{5.5}$$

i.e. W is the disjunction into subpaths of an o–d-path with respect to the threat zones and c_i are the related threat parameters.

Proposition 4.1. *Problem 5.4 and Problem 5.5 are equivalent in the following sense: If W is an optimal o–d-path for Problem 5.4 it is also optimal for Problem 5.5 and vice versa.*

Note that minimizing the probability of more than $k \geq 1$ detections is equivalent to maximizing the probability of less than or equal to k detections. This results in a problem of finding an o–d-path with minimum weighted length:

$$\text{maximize} \quad P\{X(W) \leq k\} \overset{\text{(Equation 5.3)}}{\Longleftrightarrow} \quad \text{maximize } g(\theta) := e^{-\theta} \left(\sum_{j=0}^{k} \frac{1}{j!} \theta^j \right)$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} g(\theta) = -\frac{1}{k!} e^{-\theta} \theta^k \overset{!}{=} 0 \overset{k>0}{\Rightarrow} \theta = 0 \text{ or } \theta = \infty$$

The only maximum in the range of $\theta \geq 0$ is obtained for $\theta = 0$. Further, since $g(\theta)$ is a decreasing function for increasing $\theta \geq 0$ we maximize the probability of less than or equal to k detections by minimizing θ. We deduce a general result for the structure of paths that maximize the probability of zero detections, the so-called probability of escape.

Lemma 4.2. *Assume that we are given a set of g closed, bounded, pathwise connected sets. The interior of each of these sets related with a threat parameter* $\alpha_m \in \mathbf{R}^+, m = 1, \ldots, g$, *represents a threat zone* TZ_m, $m = 1, \ldots, g$. *The threat parameter outside the threat zones is* $\alpha_0 \in \mathbf{R}^+$. *The threat parameter on the boundaries is the minimum of the threat parameters of the adjacent regions. Let o–d-path W be an optimal solution of Problem 5.4. Then W consists of linear segments and/or boundary segments of the threat zones and the break points between such segments are located only on boundaries of threat zones.*

Proof. According to Lemma 4.1 finding an o–d-path minimizing the risk of detection is equivalent to determining an o–d-path with minimum weighted length. Let o–d-path W be an optimal solution of Problem 5.4. Hence W is optimal for Problem 5.5. Since W is an o–d-path, it can be subdivided into subpaths with respect to the different threat zones it passes through (see Figure 5.2). Along each subpath the threat parameter is constant. The points in which the parameters switch are located on the boundaries of the threat zones. Consider two successive switch points b_k and b_{k+1} and the subpath W_k of the optimal path W connecting both. Because of the principle of optimality for shortest paths problems W_k has to be a b_k–b_{k+1}-path with minimum (weighted) length. Since we assumed that the threat zones are pathwise connected and using the characteristic of switching points, W_k is included in the closure of exactly one threat zone $\overline{TZ_k}$. Hence, the problem of finding the optimal subpath W_k is equivalent to finding the shortest path in the presence of barriers between b_k and b_{k+1}, where the barrier set is $B = \mathbb{R}^2 \setminus TZ_k$. According to Smith (1974) the shortest Euclidean path between two points in the plane in the presence of barriers consists of line segments and of pieces of arcs (or isolated points) along the boundary of the barriers. So the subpath W_k consists only of line segments and boundary segments of B and has break points only on boundaries of B. Since W_k was chosen arbitrarily this holds for each of the subpaths of W. The switching points are break points between the subpaths and hence W consists only of line segments and boundary segments with break points only on boundaries of threat zones. □

Observe that this result can be generalized to any metric d induced by a twice differentiable norm (cf. Pfeiffer 2006).

The problem to determine an o–d-path with minimum weighted length depends vitally on the shape and structure of the threat zones. In order to address different shapes and also threat levels depending from the distance to threat sources we model the threat zones as nested convex polygons. Threat zones are described by the interior of convex polygonal sets which can be nested into each other. That means in a convex set there can be located another convex polygon and so on (see Figure 5.3). Formally, the

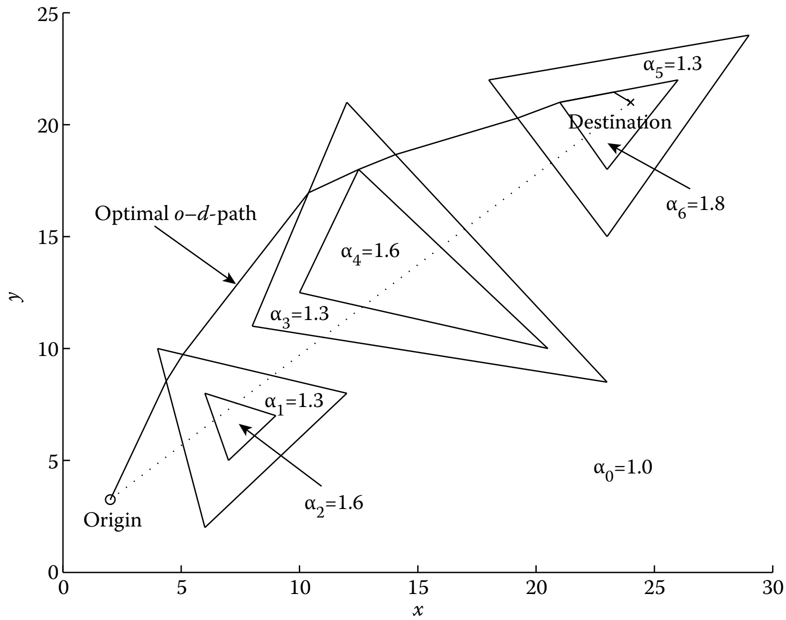


FIGURE 5.3 Optimal solution of Problem 5.4 for nested convex polygonal threat zones.

boundary of every region TZ_m, $m=1, \ldots, g$, is given by a finite set of closed, simple, pairwise nonintersecting, polygonal curves $\gamma_{m\pi}$, $\pi \in \Pi_m$, with a finite set of extreme points, where Π_m is the respective index set. Hence:

$$\partial(TZ_m) = \bigcup_{\pi \in \Pi_m} \gamma_{m\pi}, \; m=1, \ldots, g.$$

The threat parameter within one zone is constant, but the dependence of the distance from a threat source can be modeled by the different zones.

In this case Lemma 4.2 turns into Lemma 4.3 given below.

Lemma 4.3. *Given the instance described in Lemma 4.2 and, additionally, that the threat zones are modeled by nested convex polygons, let o–d-path W be an optimal solution of Problem 5.4. Then W is a piecewise linear path with break points only on boundaries of the threat zones.*

Proof. Follows from Lemma 4.2, the fact that in the case of polygons the boundary segments are linear and Lemma 4.1. □

Problem 5.5 is a continuous shortest path problem with weighted Euclidean distances, which is somewhat difficult to solve for larger problems. Among others, this problem has been investigated by Mitchell and Papadimitriou (1991), Mata and Mitchell (1997) and Reif and Sun (2001). It is related to problems investigated in robotic motion planning. Mitchell and Papadimitriou (1991) developed an algorithm which allows an ε-optimal path to be traced between a given source and any arbitrary point in the plane. The algorithm exploits the fact that shortest paths obey Snellís Law of Refraction at region boundaries, a local optimal property of shortest paths that is well known from the analogous optics model.

An example problem with three triangular nested threat zones is illustrated in Figure 5.3. The threat parameter outside the threat zones and along their boundaries is $\alpha_0=1$ and the threat parameters within the interior of the threat zones are specified in the figure. For this instance, Problem 5.4 was solved numerically using Matlab 7.0.4. The o–d-path minimizing the risk of being detected is depicted. Naturally, the shape of the path depends on the threat parameters.

5.5 Flying Time versus Risk Tradeoff

In the second model the utility of the target is time sensitive, because say after a certain duration of time it moves away into cover or if a sequence of targets needs to be visited. In this case there exists a time criticality for the mission execution. Thus the decision-maker cannot exclusively focus on finding a path which minimizes the number of detections but has also to consider the related flying time.

The problem is to minimize simultaneously the risk of being detected more than k times as well as the flying time $T(W)$ of the UAV depending on the chosen path W. These competing tasks yield a bicriteria optimization problem:

$$\begin{aligned} &\text{minimize}\, f_1(W) = P\{X(W) > k\} \\ &\text{minimize}\, f_2(W) = T(W) \\ &\text{subject to } W \text{ is an } o\text{–}d\text{-path} \end{aligned} \tag{5.6}$$

Given a transportation network, Hallam et al. (2001) modeled this problem for an application in submarine routing as a bicriteria shortest path problem on the network with additional time restrictions. We will discuss the continuous case.

For the analysis of Problem 5.6, some notation from multicriteria optimization is needed which is presented using a problem specific notation. For general results and a detailed survey of multicriteria optimization (see, for example, Ehrgott 2000).

Let $\mathcal{W}$ be the set of all possible o–d-paths and let $f(W) = (f_1(W), f_2(W))$ be any feasible objective vector of Problem 5.6. A solution $W^\star \in \mathcal{W}$ is called *Pareto optimal*, if there is no $W \in \mathcal{W}$ such that $f(W) < f(W^*)$, i.e. $f_j(W) \le f_j(W^*)$, $j = 1, 2$, and $f(W) \ne f(W^*)$. If W^* is Pareto optimal, $f(W^*)$ is called *efficient*. The set of all Pareto optimal solutions $W^\star \in \mathcal{W}$ is denoted W_{par}, the Pareto set. The set of all efficient points $z = f(W^*)$ $\in f(\mathcal{W})$, where $W^* \in W_{\text{Par}}$, is denoted Z_{eff}, the efficient set. A solution $W^\star \in \mathcal{W}$ is called *weakly Pareto optimal* if there is no $W \in \mathcal{W}$ such that $f(W) \ll f(W^*)$, i.e. $f_j(W) < f_j(W^*)$, j = 1, 2. The point $z^* = f(W^*)$ is then called *weakly efficient*. The weakly Pareto optimal and weakly efficient sets are denoted $X_{\text{w-Par}}$ and $Z_{\text{w-eff}}$, respectively.

Observe that the application of a weighted sum or Lagrangian approach to the bicriteria Problem 5.6 does not suffice to generate the efficient set, since it is nonconvex. In order to determine its efficient set we utilize the ε-constraint method. The method and related theoretical results are described in Chankong and Haimes (1983) Its advantage is that only one of the original objectives is minimized, while the other is transformed into a constraint. For j ∈ {1, 2} the original problem is substituted by the ε-constraint problem.

$$\begin{aligned} &\text{minimize fj(W)} \\ &\text{subject to } f_m(W) \le \varepsilon_{m,}\ \text{m} \in \{1,2\} \backslash \{j\} \qquad\qquad P_j(\varepsilon) \\ &\qquad\qquad W \in W, \end{aligned}$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2) \in \mathbf{R}^2$.

According to Chankong and Haimes (1983) the following interrelation between the solutions of $P_j(\varepsilon)$ and the solutions of the bicriteria Problem 5.6 holds. Given an optimal solution $\hat{W}$ of $P_j(\varepsilon)$ for some $j \in \{1, 2\}$, then $\hat{W}$ is weakly Pareto optimal for Problem 5.6. Furthermore, the solution $\hat{W} \in W$ is Pareto optimal for Problem 5.6 if and only if there exists an $\varepsilon \in \mathbf{R}^2$ such that $\hat{W}$ is an optimal solution of $P_j(\varepsilon)$ for all $j = 1, 2$.

In our case the two possible ε-constraint problems represent two special problems which are of great interest on their own. $P_1(\varepsilon)$ yields the problem of minimizing the risk of being detected more than k times while there is given a time period $\varepsilon_2 = T$ within which the mission has to be completed:

$$\begin{aligned} &\text{minimize } P\{X(W) > k\} \\ &\text{subject to } T(W) \leq T \qquad\qquad P_1(\varepsilon) \\ &W \in W. \end{aligned}$$

On the other hand, $P_2(\varepsilon)$ describes the problem of finding a path with minimum flying time of the UAV under the restriction that a certain risk $\varepsilon_1 = R$ of being detected more than k times is not exceeded:

$$\begin{aligned} &\text{minimize } T(W) \\ &\text{subject to } P\{X(W) > k\} \leq R \qquad\qquad P_2(\varepsilon) \\ &W \in W. \end{aligned}$$

To minimize the flying time is, for example, relevant if one is interested in the routing or dispatching of a UAV to a number of targets (see Mishra et al. 2004).

We generate efficient solutions of Equation 5.6 by solving both problems for different values for T and R. For this purpose, we first reformulate the corresponding problems. In Section 5. 4 we showed that to minimize $P\{X(W) > k\}$ is equivalent to minimize the sum of the weighted length of the subpaths $\theta = \sum_{i=1}^{n} c_i l(W_i)$. According to Equation 5.3 $P\{X(W) > k\}$ is a function of θ and so each risk restriction R corresponds to a restriction on the weighted length denoted θ_R. We assume that the UAV moves with constant speed s. Hence, since $T(W) \cdot s = l(W)$, to minimize the (unweighted) length $l(W) = \sum_{i=1}^{n} l(W_i)$ of the o–d-path is equivalent to minimize the flying time $T(W)$. According to Lemma 4.3 an optimal o–d-path through nested convex polygonal threat zones is piecewise linear with break points only on the boundaries of the threat zones. Denote $(x_0, y_0) := (x_o, y_o), (x_n, y_n) := (x_d, y_d)$ and (x_i, y_i), $i = 1, \ldots, n-1$, as the unknown break points of the o–d-path W located on the boundaries of the threat zones. Then:

$$l_i := l(W_i) = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}.$$

The boundaries of the threat zones are given by a finite number of linear equations with fixed ranges for the coordinates. If we know on which boundary the break points are located, we can eliminate the y-variables by using:

$$y_i = \frac{e_i - a_i x_i}{b_i}, \quad i = 1, \ldots, n-1,$$

$$r_i \leq x_i \leq u_i, \; i = 1, \ldots, n-1$$

where the respective facet of the polygonal threat zone is described by the parameters e_i, a_i, b_i, r_i and u_i. Under this presumption problems $P_1(\varepsilon)$ and $P_2(\varepsilon)$ can be transformed into:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} c_i l_i \\ &\text{subject to} && \sum_{i=1}^{n} l_i \leq T \cdot s && P'_1(\varepsilon) \\ &&& r_i \leq x_i \leq u_i, \quad i = 1,\ldots,n-1 \\ &&& x \in \mathbb{R}^{n+1} \end{aligned}$$

and

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} l_i \\ &\text{subject to} && \sum_{i=1}^{n} c_i l_i \leq \theta_R && P'_2(\varepsilon) \\ &&& r_i \leq x_i \leq u_i, \quad i = 1,\ldots,n-1 \\ &&& x \in \mathbb{R}^{n+1} \end{aligned}$$

As long as the sequence of edges of the threat zones that are intersected by the path is unchanged, both problems are convex optimization problems since both the objective functions and the constraint functions are convex and they are defined on an open region. But since the sequence of edges of the optimal path is initially unknown and its determination is part of the optimization problem, the original problems $P_1(\varepsilon)$ and $P_2(\varepsilon)$ are in general nonconvex optimization problems.

As a numerical example we consider again the instance introduced in Section 5.4 in Figure 5.3. We solve both problems $P_1(\varepsilon)$ and $P_2(\varepsilon)$ for varying values for the time constraint T and the risk constraint R, respectively. This is realized by solving for all feasible edge sequences the respective problems $P_1'(\varepsilon)$ and $P_2'(\varepsilon)$ with Matlab 7.0.4. The speed of the UAV is assumed to be constant, i.e. $s=1$. In Figure 5.4 several optimal paths are depicted which minimize the risk of being detected without exceeding the respective prescribed limit of flying time. Obviously, there is a lower bound on the flying time, namely the time needed for the straight line connection between the origin and the destination. Relaxing the time constraint results in paths more and more evading, even meandering ($T=28.7$) through the zones with increased threat. Observe that there is an upper limit for the time restriction (here: $T=30.7$). Even if we allow for time values above it the structure of the optimal path remains unchanged. We find that in this case the optimal paths coincide with the optimal solution for the unrestricted problem discussed in Figure 5.3. This demonstrates the coherence of our problem formulation and again, it points out that eschewing the threat zones does not further decrease the risk.

For the same example problem, paths minimizing the flying time while observing the respective prescribed limit of risk are illustrated in Figure 5.5. Note that for the given example the straight and therefore fastest connection between the origin and the destination is related to the highest risk of detection ($\theta_R=36.2$). Just like in the converse problem there exists a lower bound for the risk ($\theta_R=34.4$), such that allowing for even longer flying times does not decrease the achievable risk any further.

Combining the results of these two ε-constraint problems yields a set of weakly efficient points (see Figure 5.6) belonging to the weakly efficient set of Problem 5.6. The weakly efficient set is characterized by the different sequences of edges of the threat zones that feasible o–d-paths cross. If one considers the problem of finding an optimal path restricted to only one of these edge sequences, the resulting optimization problem is convex. But since for varying time and risk restrictions, respectively, different sequences lead to optimal solutions, the overall problem is nonconvex as can be seen in Figure 5.6. The crosses indicate dominated solutions while circles represent weakly efficient solutions. Consequently,

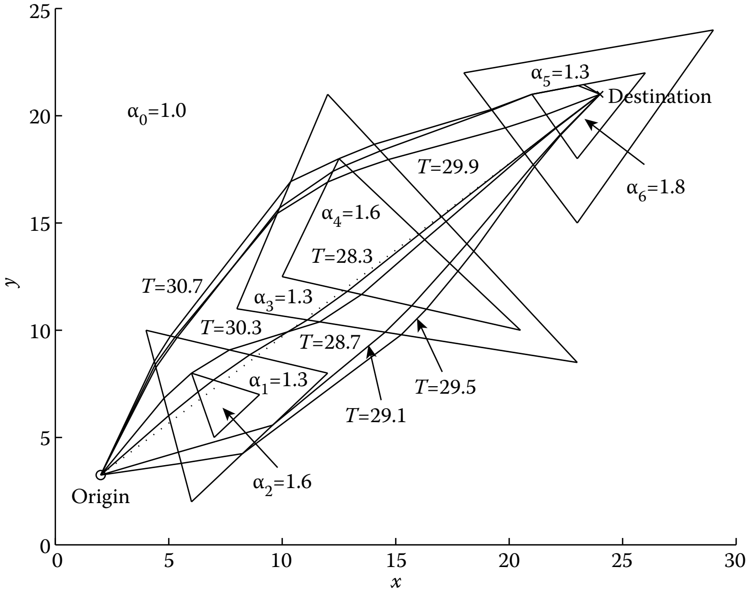


FIGURE 5.4 Example problem for flight time/risk tradeoff study.

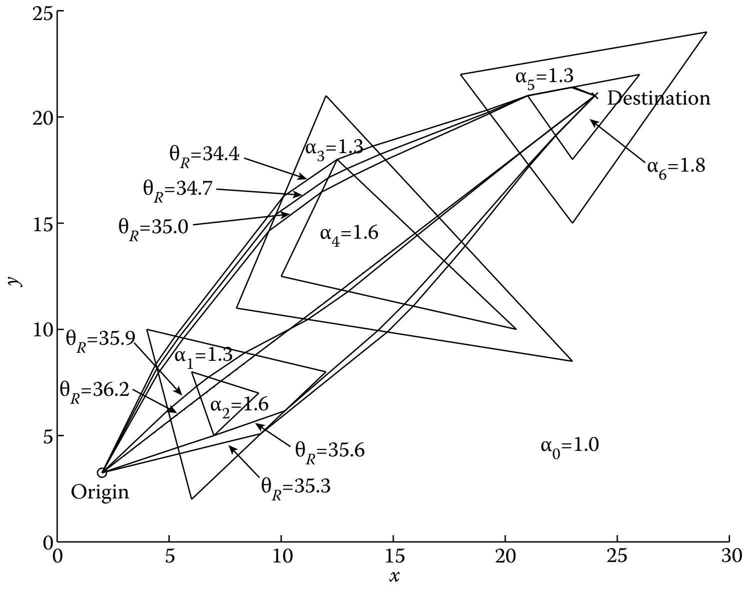


FIGURE 5.5 Optimal *o*–*d*-paths for the example problem.

each of the weakly efficient points in Figure 5.6 corresponds to a weakly efficient path. Some of them are depicted in Figures 5.4 and 5.5.

It is characteristic for bicriteria problems that a relaxation of the bound on one of the objectives results in an improvement of the other objective. In our case, the tighter the time restriction, the more risk one has to accept and vice versa. It is up to the decision maker to choose a most preferred path, and the efficient set provides useful tradeoff information between flying time and risk to support the decision process.

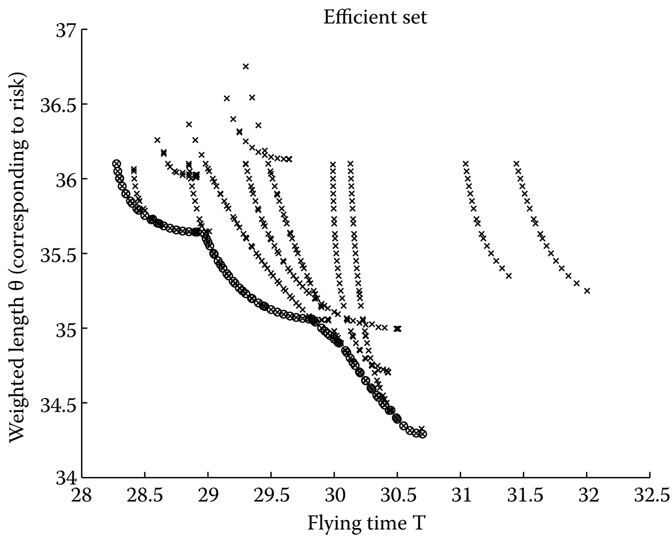


FIGURE 5.6 Efficient set of Problem 5.6.

5.6 Modeling Enhancements

In situations where several paths are associated with the same escape probability, time until first detection can be regarded as a tie-breaking rule—our first enhancement investigates this situation. Our second enhancement considers the situation where the first detection of the UAV entails an fatal attack with a certain delay. This leads to consideration of UAV sacrifice when the target is close by.

5.6.1 Expected Time Till First Detection

In Section 5.4, we solved the problem of finding paths that maximize the probability of escape. There might be environmental structures such that the optimal solution is not unique and therefore several paths exist with the same minimum probability of detection (see, for example, Figure 5.7). Then the operator needs an additional decision criterion in order to choose the safest path. The probability of escape is the same on these paths, but assuming that at least one detection occurs, that path will be preferred on which the first detection arrives as late as possible. For such cases we examine the secondary objective of expected time till first detection as a tie breaker. We assume $s=1$ and hence time is equivalent to distance. Let T_1 be the random variable that measures the time till the first detection on the o–d-path W, let $W=\biguplus_{i=1}^{n} W_i$ be a decomposition of W into n subpaths W_i, let $l_i=l(W_i)$ be the length of subpath W_i, and c_i the constant threat level on W_i.

A is the event that one (or more) detections occur on W, and A_i, $i=1, \ldots, n$, is the event that the first detection on subpath occurs on W_i. Since the events $A_i, i=1, \ldots, n$, are mutually exclusive we have $A=\biguplus_{i=1}^{n} A_i$. Therefore the conditional expectation for the time until the first detection T_1 given at least one detection is given by:

$$E[T_1 \mid A]=\frac{\sum_{i=1}^{n} E[T_1 \mid A_i]\cdot P(A_i)}{P(A)},$$

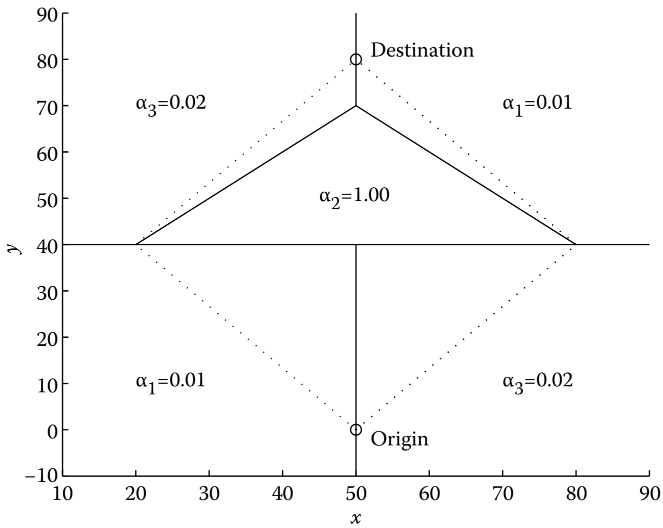


FIGURE 5.7 Two optimal paths with the same probability of escape.

where $P(A)=1-P\{X(W)=0\}=1-\exp\left(-\sum_{i=1}^{n} c_i l_i\right)$ and $P(A_i)=\exp\left(-\sum_{j=1}^{i-1} c_j l_j\right)(1-e^{-c_i l_i})$.

Under the condition that the UAV is detected on subpath W_i for the first time it passes the (i–1) preceding subpaths without being detected and it follows that:

$$E[T_1 \mid A_i]=\sum_{r=1}^{i-1} l_r + E[Q_i \mid C],$$

where Q_i is the random variable that measures the time flown along W_i until the first detection and C is the event one (or more) detection on W_i. For the events C_k, $k \geq 1$, that exactly k detections occur on W_i, it holds $C=\biguplus_{k=1}^{\infty} C_k$ and therefore:

$$E[Q_i \mid C]=\frac{\sum_{k=1}^{\infty} E[Q_i \mid C_k]\cdot P(C_k)}{P(C)},$$

where $P(C)=1-e^{-c_i l_i}$, $P(C_k)=\frac{1}{k!}(c_i l_i)^k e^{-c_i l_i}$ and $E[Q_i \mid C_k]=\frac{l_i}{k+1}$, because given k arrivals in a homogeneous Poisson process on an interval these are spread uniformly over the interval (cf, Larson and Odoni 1981). Thus the conditional expectation on the subpath simplifies to:

$$E[Q_i \mid C]=\frac{e^{c_i l_i}-c_i l_i-1}{c_i(e^{c_i l_i}-1)}.$$

Summarizing the above results leads to the following expression for the conditional expected time until the first detection, given at least one detection occurs:

$$E[T_1 \mid A] = \frac{\sum_{i=1}^{n}\left(\left(\sum_{r=1}^{i-1} l_r + \frac{e^{c_i l_i} - c_i l_i - 1}{c_i\left(e^{c_i l_i} - 1\right)}\right)\cdot\left(1 - e^{-c_i l_i}\right)e^{-\sum_{j=1}^{i-1} c_j l_j}\right)}{\left(1 - e^{-\sum_{i=1}^{n} c_i l_i}\right)} \tag{5.7}$$

For the example depicted in Figure 5.7 the probability of escape for the two optimal paths is 0.223. For the right path the expected time until the first detection given at least one detection is 30.6, and for the left path the value is 46.6. If the operator uses this criterion as tie breaker, he will choose the left path. The results imply that there is not only a tradeoff between the probability of escape and the length of the path, but that there is also an issue of when the risk should be undertaken. At an intuitive level this makes logical sense, since taking a risk at the beginning of the path should be significantly worse than accepting a risk towards the end of the path, when the UAV is near the destination.

5.6.2 Delay Between Detection and Fatal Attack

Another objective arises if the decision maker differentiates between risk for the UAV and risk for the mission. This could be modeled by presuming that there is a delay between the first detection of the UAV and a fatal attack which destroys the UAV and aborts the mission. That means if the UAV manages to reach the target despite being detected, but before being destroyed, i.e. in the time gap which exists due to the delay until the fatal attack, the mission could be completed. So in this model we are only concerned about the first detection of the UAV. We present two approaches, one presumes a constant delay between detection and fatal attack whereas the other deals with a randomly distributed delay.

5.6.2.1 Constant Delay

In this section, we assume that there is a constant time delay D between the first detection of the UAV and a fatal attack. That means that the mission still can be completed if the UAV manages to reach the destination point in the time D after the first detection. The aim is to find a path, which guarantees best the success of the mission under these terms.

We assume a constant speed $s=1$. Then all points from which the UAV can get to the destination point d within time D are contained in the circle $B(d, D)$ with center d and radius D (see Figure 5.8).

The problem turns into the task to route the UAV with minimum risk of being detected between the origin o and the boundary of the target circle $\partial B(d, D)$. This is described by the following optimization problem

$$\begin{aligned} &\text{minimize} && P\{X(W)>0\} \\ &\text{subject to} && W \text{ is an } o-z\text{-path} \\ & && \text{where } z \in \partial B(d,D). \end{aligned} \tag{5.8}$$

According to Equation 5.3 and Lemma 4.1 this is equivalent to:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} c_i l(W_i) \\ &\text{subject to} && W = \biguplus_{i=1}^{n} W_i \end{aligned}$$

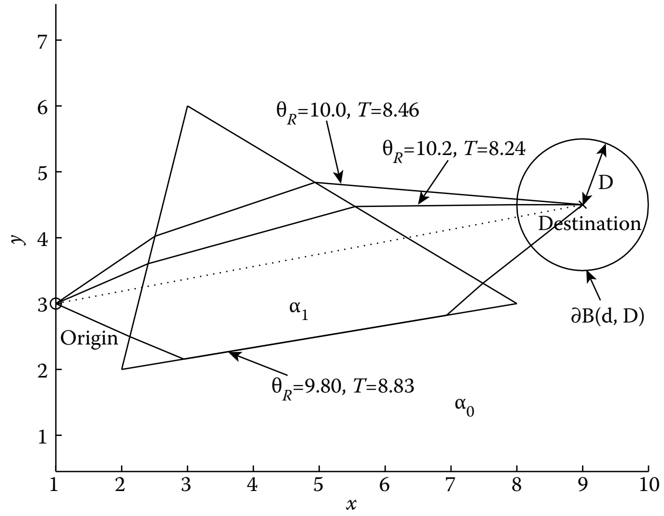


FIGURE 5.8 Weakly efficient paths under the assumption of a constant time delay D between first detection and mission abortion ($s=1$).

i.e. is the disjunction into subpaths of an o–z-path with respect to the threat zones and c_i are the related threat parameters.

Note that the distribution of threat parameters inside the circle $B(d,$ D) has no impact on the success of the mission and hence on the structure of an optimal path. Moreover, if the interior of the circle int $B(d, D)$ is related to exactly one threat parameter, an optimal solution of Problem 5.8 corresponds to an optimal solution of Problem 5.4 and vice versa.

Lemma 6.1 *Consider the same situation as in Lemma 4.2 and additionally a constant time delay D between the first detection and a fatal attack. We assume that the interior of the circle int B(d,D) is related to exactly one threat parameter* α_j, $j \in \{0, \ldots, g\}$ *and constant speed* $s=1$. *Then a path which is optimal for Problem 5.4 is also optimal for Problem 5.8 and vice versa.*

Proof. The optimal solution of Problem 5.8 is according to Lemma 4.1 an o–z-path with minimum weighted length. Since $\min_{z \in \partial B(d,D)} \alpha_j l(z, d) = \alpha_j D$, it corresponds to an o–z–d-path with minimum weighted length. Assume that x_{n-1} is the last break point of an optimal solution of Equation 5.8 on an edge of a threat zone before reaching the point $z \in \partial B(d, D)$.

Proposition: The weighted length of the subpath between x_{n-1} and d equals the weighted length of the straight line segment $\overline{x_{n-1}d}$.

Proof of proposition: Suppose that the weighted length of the subpath between x_{n-1} and d is not equal to the weighted length of the straight line segment $\overline{x_{n-1}d}$. The principle of optimality implies that it is, however, an optimal path from x_{n-1} to d. Let α_k be the unique threat parameter related to the subpath between x_{n-1} and the boundary of the circle $\partial B(d, D)$. Let p be the intersection point of the straight line segment that connects x_{n-1} and the destination d and the boundary of the circle, i.e. $p = \overline{x_{n-1}d} \cap \partial(B(d, D))$. Then it holds $p = \overline{x_{n-1}d} \cap \partial B(d, D)$. Then:

$$\alpha_k l(x_{n-1}, z) + \alpha_j l(z, d) < \alpha_k l(x_{n-1}, p) + \alpha_j l(p, d)$$

Since $l(z, d) = l(p, d)$ this leads to:

$$l(x_{n-1}, z) + l(z, d) < l(x_{n-1}, d)$$

This is contradiction to the triangle inequality, since l is a metric. As a result $\overline{x_{n-1}d}$ is the optimal length of the subpath from x_{n-1} to d.

Therefore the number of break points on an optimal o–z–d-path is the same as for a corresponding optimal o–d-path, and for a given setup, optimal solutions of Equation 5.8 are also optimal for Equation 8.4 and vice versa. □

Again we can consider the bicriteria problem of minimizing risk and flying time simultaneously. According to the discussion above the weakly efficient paths for the problem with constant time delay coincide with those of Problem 5.6. For the example given in Figure 5.8 several weakly efficient paths are depicted.

5.6.2.2 Randomly Distributed Delay

If the delay between the first detection of the UAV is not constant, but itself a random variable, a different model is required. We introduce the random variable Z, which measures the time until a fatal attack under the assumption that the UAV is detected on its o–d-path. Thus, Z is the sum of the time until the first detection T_1 under the condition of the event A given at least one detection and the delay between the first detection and the fatal attack D. Hence the expected value of Z is:

$$E[Z]=E[T_1|A]+E[D],$$

where the first term is given by Equation 5.7 and the second depends on the probability distribution describing the delay D. For example, the PDF for the delay D can be assumed to be proportional to $t^2e^{-\beta t}$, since this reflects the retarding moment and that after a certain peak the probability of being shot down decreases again. In addition, the delay does not depend of the location of the detection, whereas the PDF of T_1 vitally depends of the actual location of the UAV. Presuming such a PDF results in an expected value for D given by $E[D]=3/\beta$. Note that the parameter β reflects the adversarial response time.

5.7 Conclusions

This chapter investigates the problem of path planning for UAVs in a hostile environment where different threat sources cause uncertainties in the mission planning. Nested convex polygonal threat zones are considered for the problem of determining the path that maximizes the UAVís escape probability in a continuous setting. Bicriteria optimization was used to explore the tradeoff between flying time versus escape probability. Expected time till first detection was regarded as a tie-breaking objective for paths with the same escape probability. The effect of delay between detection and fatal attack was also discussed. Our conclusion is that UAV routing can be effectively modeled using probabilistic constructs. Furthermore, for the case of maximizing escape probability, the resultant optimization problems are numerically tractable for many situations. Future work should consider additional issue that might be of interest to the military analyst, e.g. closed-loop path planning when the UAV can sense that it has been detected (but not yet attacked), and coordinated UAV fleet operations.

References

Anderson, E., Beard, R., and McLain, T. 2005. Real-time dynamic trajectory smoothing for unmanned air vehicles. *IEEE Transactions on Control Systems Technology*, 13(3), 471–477.

Bortoff, S.A. 2000. Path planning for UAVs. In *Proceedings of the American Control Conference,* June 28–30, Chicago, IL, 364–368.

Casas, I., Malik, A., Delmelle, E., Karwan, M., and Batta, R. 2007. An automated network generation procedure for routing of unmanned aerial vehicles (UAVs) in a GIS environment. *Networks and Spatial Economics*, 7, 153–176.

Chankong, V. and Haimes, Y.Y. 1983. *Multiobjective Decision Making: Theory and Methodology*. Elsevier Science Publishing, NY.

Dogan, A. 2003. Probabilistic approach in path planning for UAVs. Presented at the IEEE International Symposium on Intelligent Control, October 5–8, Houston, TX.

Dogan, A. and Zengin, U. 2006. Unmanned aerial vehicle dynmaic-target pursuit by using probabilistic threat exposure map. *Journal of Guidance, Control, and Dynamics*, 29(4), 944–954.

Ehrgott, M. 2000. Multicriteria optimization. In: *Lecture Notes in Economics and Mathematical Systems 491*. Springer-Verlag, Berlin.

Erkut, E. and Ingolfsson, A. 2000. Catastrophe avoidance models for hazardous materials route planning. *Transportation Science*, 34, 165–179.

Erkut, E. and Ingolfsson, A. 2005. Transport risk models for hazardous material: revisited. *Operations Research Letters*, 33, 81–89.

Hallam, C., Harrison, K.J., and Ward, J.A. 2001. A multiobjective optimal path algorithm. *Digital Signal Processing*, 11, 133–143.

Hespanha, J.P., Kızılocak, H.H., and Ateşkan, Y.S. 2001. Probabilistic map building for aircraft-tracking radars. In *Proceedings of the American Control Conference,* June 25–27, Arlington, Virginia, Vol. 6, 4381–4386.

Jin, H., Batta, R., and Karwan, M.H. 1996. On the analysis of two new models for transporting hazardous materials. *Operations Research*, 44(5), 710–723.

Jun, M. and Dí'Andrea, R. 2002. Path planning for unmanned aerial vehicles in uncertain and adversarial environments. In S. Butenko, R. Murphey, and P. Pardalos, editors, *Cooperative Control: Models, Applications and Algorithms*. Kluwer Academic Publisher, Dordrecht, 95–111.

Larson, R.C. and Odoni, A.R. 1981. *Urban Operations Research*. Prentice Hall, Englewood Cliffs, NJ.

Mata, C.S., and Mitchell, J.S.B. 1997. A new algorithm for computing shortest paths in weighted planar subdivisions (extended abstract). In *Proceedings of the Thirteenth Annual Symposium on Computational Geometry, June 4–6, Nice, France*, ACM Press, 264–273.

McLain, T.W., Chandler, P.R., Rasmussen, S., and Pachter, M. 2001. Cooperative control of UAV rendezvous. In *Proceedings of the American Control Conference,* June 25–27, Arlington, Virginia, Vol. 3, 2309–2314.

MCWP3-42.1. 2003. Unmanned aerial vehicle operations. Marine Corps Warfight Publication, Department of the Navy, Headquarters United States Marine Corps, Washington, DC.

Mishra, S., Batta, R., and Szczerba, R.J. 2004. A rule based approach for aircraft dispatching to emerging targets. *Military Operations Research*, 9(3), 17–30.

Mitchell, J.S.B. and Papadimitriou, C.H. 1991. The weighted region problem: finding shortest paths through weighted planar subdivision. *Journal of the Association for Computing Machinery*, 38, 18–73.

Nielsen, P.D., Noor, A.K., and Venner, S.L. 2003. The next century of air power. *Mechanical Engineering*, 125, 34–40.

Pfeiffer, B. 2006. Path Planning and Location Problems with Varying Environments. Shaker-Verlag, Aachen.

Reif, J.H. and Sun, Z. 2001. An efficient approximation algorithm for weighted region shortest path problem. In B. Donald, K. Lynch, and D. Rus, editors, *New Directions in Algorithmic and Computational Robotics*. A. K. Peters, Boston, MA, 191–203.

Smith, D.R. 1974. *Variational Methods in Optimization*. Prentice-Hall, Englewood Cliffs, NJ.

6

Modeling the End-to-End Military Transportation Problem

James T. Moore
Air Force Institute of Technology

J. Wesley Barnes
The University of Texas at Austin

Raymond R. Hill
Air Force Institute of Technology

6.1 Introduction

Moving people, equipment, and supplies to support military endeavors has been accomplished since the dawn of recorded history. Whenever people in distant lands have been engaged in conflicts with each other, the ability to move people (warriors) and their equipment and supplies successfully has been a key to the outcome of the conflict. The opponent who has to move the furthest to engage in conflict has found the accomplishment of the deployment and distribution task critical. In order to succeed, the "right things me must be at the right place at the right time". Historical records are filled with examples of where the winner has done this successfully and the loser has not.

A few notable examples are offered to support this observation. Alexander the Great conquered a large portion of the known world by moving large number of soldiers across large distances, living off the land, and incorporating conquered people into his army (thus extending his support and supply lines). Hannibal moved armies and their tools of war over mountains to attack Rome. Napoleon's armies were able to cover large distances and secure large holdings because of their ability to carry supplies and supplement their supplies from the conquered economies. Skipping ahead to more recent times, Hitler's armies moved at lightening speed because of their superior support mechanism. Yet, when these vast armies invaded Russia, they could not win. One of the major reasons for failure was their lack of adequate winter clothing.

Today, the US military supports a far flung network of people and installations. American troops are located around the world and many are engaged in hostile actions. US military installations and embassies circle the globe, and each requires support and supply deliveries. At this writing, the US military is in the process of establishing a new command, the African command, which will require a significant level of long distance support.

The mission of moving people, equipment, and supplies around the world is vast and immensely important. In order to study and model this problem, USAF Air Mobility Command (AMC), the University of Texas at Austin (UT), and the Air Force Institute of Technology (AFIT) formed the AMC-UT-AFIT Research Consortium. The Consortium research effort was funded by the Air Force Office of Scientific Research (AFOSR). Rather than try to model the whole deployment and distribution system, specific aspects of the system have been modeled and are presented in this chapter. The aspects of the system we investigated and modeled are pallet packing, aircraft loading, strategic airlift, aerial refueling, aircrew scheduling, and theater distribution. Before delving into these modeling efforts individually, we briefly consider the foundations of the modeling efforts which include heuristics, abstract algebra (in the context of group theory) and tabu search (TS).

6.2 Modeling Foundations

We will first discuss the heuristic foundations.

6.2.1 Heuristics

In practice, the time and effort associated with obtaining a pristine provably optimal solution to a combinatorial optimization problem (COP) is usually less practical, even when it is possible, than using an easily computed heuristic approach to acquire near-optimal solutions (Hill et al., 2008). As *practical* problem formulations increase in size, finding proven optimal solutions usually requires intolerable computing time and storage space. Due to such things as the imprecision of real-world problem data, models of large practical problems are invariably abstractions of reality possessing fragile optimal solutions that most often have little practical value.

6.2.2 Group Theory

Since group theory has been very effectively used in solving several of the transportation problems addressed by the Consortium (Wiley, 2001; Crino, 2002; Crino et al., 2004; Barnes et al., 2004; Burks, 2006; Burks et al., 2008), a brief conceptual overview of group theory is now presented.

Abstract groups are simply explained as sets of objects, together with a method of combining its elements that is subject to a few simple rules (Colletti, 1999; Baumslag and Chandler, 1968).

A *semigroup* is a nonempty set G together with a fixed binary operation $\oplus$ that satisfies the following conditions:

1. $x \oplus y \in G, \forall x,y \in G$; the operation is closed.
2. $(x \oplus y) \oplus z = x \oplus (y \oplus z)\ \forall x,y,z \in G$; the operation is associative.

A *group* is a semigroup that satisfies the additional conditions:

1. There exists $e \in G \ni \forall x \in G, e \oplus x = x \oplus e = x$; there exists a unique identity.
2. For each $x \in G$, there exists $x^{-1} \in G \ni x^{-1} \oplus x = x \oplus x^{-1} = e$; there exists a unique inverse.

Groups have some elementary properties that are used in this research. These properties, presented as a theorem with a proof in Barnard and Neill (1996), are:

1. For $x, y \in G$, if $x \oplus y = e$, then $x = y^{-1}$ and $y = x^{-1}$
2. $(x \oplus y)^{-1} = y^{-1} \oplus x^{-1} \; \forall x, y \in G$
3. $(x^{-1})^{-1} = x \; \forall x \in G$
4. For $x, y, z \in G$, if $z \oplus x = z \oplus y$, then $x = y$ and if $x \oplus z = y \oplus z$ then $x = y$

A *permutation of a set A* is a function from A into A, which is both one to one, and onto (Fraleigh, 1976). The *symmetric group on n-letters,* S_n is the group of all permutations of set A if A is the finite set of "letters", {1, 2, 3, …, n}(Fraleigh, 1976). S_n contains many subgroups that obey the four above properties where S_n is formally a subgroup of itself.

Any $\pi \in S_n$ may be represented in the standard or cyclic form. The standard form uses a $2 \times n$ array comprising a one to one and onto function whose domain (top row) and image (bottom row) are the integers {1, 2, …, n}, i.e. the standard notation for any π is

1	2	3	4	5	….	n
$\pi(1)$	$\pi(2)$	$\pi(3)$	$\pi(4)$	$\pi(5)$	….	$\pi(n)$

The equivalent cyclic form is more efficient. $\pi \in S_n$ may possess any number of cycles bounded between 1 and n. The cycle $(i,j,k,l) \subset \pi$ sends i to j, j to k, k to l, and l back to i. The process continues with another cycle and iterates until all members of {1, 2, …, n} have been used. A cycle of length κ is a cycle containing κ elements (Sagan, 1991). Unit cycles, possessing one element, are implied and not explicitly written in cyclic notation. Many examples of permutations in both forms are given in Sagan (1991).

The product of two permutations, $\pi \oplus \sigma$, is obtained by the closed associative binary operator, function composition. Given $x \in S_n$, then $(\pi \oplus \sigma)(x) = \sigma(\pi(x))$. For S_n, conjugation of one permutation, p, by another, q (the *conjugator* permutation), yields a new permutation, x, while maintaining p's cyclic structure, i.e. $x = q^{-1} \oplus y \oplus q = y^q$ and each letter in y is replaced by its successor letter in q. If $x = (1,3,2)(4)(5)$, and $k = (3,4,5)$, $x^k = (1,4,2)(5)(3)$.

Groups may be partitioned into conjugacy classes. In general, any group G inherently self-partitions into mutually exclusive and exhaustive conjugacy classes defined as CClass $(G, g) = \{g^h : h \in G\}$, the set of all elements $\{h^{-1} \oplus g \oplus h : h \in G\}$ for $g \in G$.

A *group action*, $_GT$, uses a group G to partition a set T into mutually exclusive and exhaustive cells called *orbits*. For the purposes of this chapter, G is a subgroup of S_n and T is the union of one or more conjugacy classes of S_n. Indeed, the set of conjugacy classes of S_n is the result of S_n acting upon itself with the conjugation as the operator associated with the group action. The *orbits* of $_GT$ are mutually exclusive and exhaust the elements of T (Colletti, 1999).

6.2.3 Basic Tabu Search

TS is a *metaheuristic*, a master strategy that forces a local heuristic search procedure to explore the solution space, X, the set of all possible solutions, beyond a local optimum (Glover and Laguna, 1997). The fundamental philosophies of TS are the following:

1. A memory structure should be used to create a robust search methodology, unlike simulated annealing and genetic algorithms that use randomness to guide the search process.
2. The solution space should be explored intelligently, i.e. the search must respond appropriately to what is occurring or has occurred during the search process.

In the following, we describe the basic components present in *any* TS. This basic TS is often so effective that it suffices for the problem under investigation. Unfortunately, the basic TS approach fails for more complex problems and the implementation of more advanced, readily available concepts is required.

First, create a solution structure which precisely defines any $x \in X$. The basic TS starts at a constructed initial incumbent solution. In the context of TS applied to a *real* COP, a *neighborhood*, N , consisting of *neighboring solutions* reachable in one step from the incumbent solution must be defined for each and every $x \in X$. Neighborhood are often defined by rule relative to the current incumbent solution. Given an objective function $f: X \rightarrow \mathbf{R}$, the triple $£=(X, f, N)$ forms the *landscape* for the search (Barnes et al. 2003). At each iteration, TS *moves* to a neighboring solution.

For clarification, consider the following simple example. Suppose, for a 3-crew/3-flight problem, the incumbent solution, x, has crew 1 assigned flight 1, crew 2 assigned flight 2, and crew 3 assigned flight 3. A *swap move*, where two crews exchange flights, could be used to define the neighborhood for x. Defining $x=(1,2,3)$ implies $N(x)=\{(2,1,3), (3,2,1), (1,3,2)\}$.

Next, the neighboring solutions are evaluated and one member is selected as the next incumbent solution in accordance with f and a tabu memory structure. Different selection rules can be used, i.e. select the first improving neighboring solution or select the best neighboring solution.

In choosing the new incumbent solution, the tabu memory structure must also be considered. The tabu memory structure can be extensive containing adaptive and dynamic components that control the search. At a minimum, a tabu memory structure provides the capability to escape from a local optimum and progress into other regions of the solution space. It achieves this by recording solution attributes that have changed during a specified number of iterations, the *tabu tenure*. Suppose that (2,1,3) was the best neighboring solution in our above example and becomes the new incumbent solution. Since crews 1 and 2 have changed positions (our defined attributes of the solution), one tabu restriction could be to forbid crews 1 and 2 from swapping again for tabu tenure iterations. Alternately, a more stringent restriction would be to forbid *both* crews 1 and 2 from being reassigned for tabu tenure iterations. Either of these restrictions would prohibit return to (1,2,3) for tabu tenure iterations, i.e. (1,2,3) is tabu! Of course, other solutions are also forbidden by the second restriction. The choice of a proper tabu memory structure can dramatically affect the success of the search. The tabu memory structure implemented is often problem specific and is an area of current research.

Aspiration criteria provide a flexible tool to overcome the restrictions imposed by a tabu memory structure. A commonly used aspiration is to allow a forbidden solution if it is superior to the best solution found so far. Clearly such a solution has not been visited recently and should be exploited!

This section concludes by demonstrating an iteration of the basic TS, as implemented in OpenTS, the Java™-based software. Figure 6.1 below displays an iteration of the OpenTS framework (Harder, 2002; Harder et al. 2004).

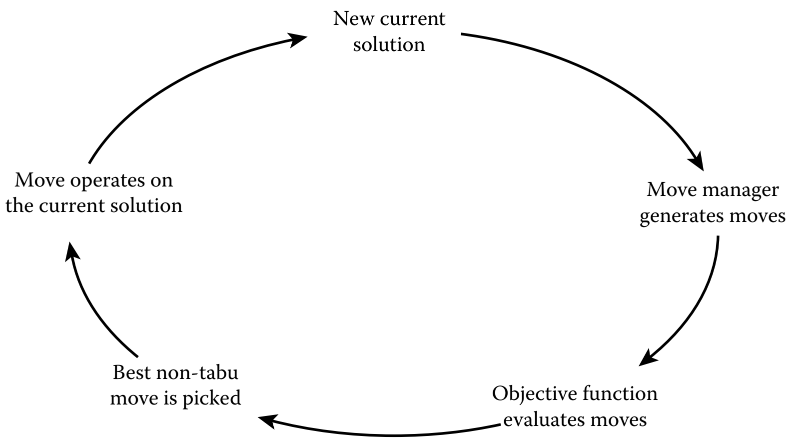


FIGURE 6.1 One iteration of OpenTS (Harde, 2002).

OpenTS starts from a user-defined initial solution. OpenTS builds a set of moves, M, which when invoked, generate the neighboring solutions and the associated f values using user-defined evaluation methods. Finally, the new incumbent solution is chosen and the next iteration begins.

Since the limited basic TS approach will almost always be insufficient for problems of practical size and complexity, let us briefly consider two somewhat more advanced TS concepts that have proved to be essential in the solution of practical problems, intensification and diversification. Intensification strategies modify the search to stay in or return to attractive regions of X where good solutions reside. For example, if the search has encountered several improving solutions in recent iterations, the tabu tenure could be adaptively reduced allowing the search to remain close by. Another intensification strategy could be invoked when the search has stagnated, i.e. no globally improving solutions have been encountered recently. In this case, the search could be started at a previously found, and stored, elite solution with a reinitialized tabu memory structure in hopes that a new search trajectory will result in nearby superior solutions.

If the search has failed to find improving solutions recently, it could be time to diversify the search. A simple diversification strategy adaptively increases the tabu tenure, forbidding return to previously visited solutions for longer periods and encouraging the search to examine unvisited regions of the solution space where the solutions differ significantly from those found previously. Another diversification approach is to record solutions recently visited in a long-term memory structure and to use that record to determine whether or not the search is trapped in a subset of X. This condition is often signaled by the repetition of an observed set of solutions exceeding a specified criterion. When this case occurs, an *escape strategy* can be invoked which could be as simple as restarting the search and as complex as redefining the goal of the search for a certain number of iterations. One such strategy would be to choose the worst neighboring solution for a sufficient number of iterations to assure that a different region of X has been achieved.

This section is, in no way, intended to be more than introductory and illustrative of the many possibilities that exist in the theory and application of TS. In the following sections, the TS associated research of the AMC-UT-AFIT Consortium directed at transportation problems within the end-to-end mobility problem will be discussed. As part of that discussion, several of the more advanced TS techniques will be described in detail.

6.3 TS Approaches to Transportation Problems within the End-to-End Mobility Problem

The Consortium has successfully attacked and solved several of the transportation problems within the end-to-end mobility problem. Concurrent with the Consortium's earlier work, a detailed review of the strategic mobility models was performed by McKenzie and Barnes (2004). We begin with the 2D bin-packing problem (Harwig, 2003; Harwig et al. 2006). This leads to the 3D pallet packing problem (Baltacioğlu, 2001; Baltacioğlu et al. 2006). We then consider the methodology developed to efficiently and effectively load pallets onto cargo aircraft (Roesener, 2006; Roesener et al. 2008) and to assist military aircraft to accomplish their ongoing resupply mission by solving the strategic airlift problem (SAP) (Lambert, 2004; Lambert et al. 2007). Early in the deployment process in the initial stages of a conflict, the aerial fleet refueling problem (AFRP) must be solved (Wiley 2001; Barnes et al. 2004), i.e. the smaller combat aircraft must fly to the region of the combat theater. Since they cannot make the flight over the ocean on a single tank of gas, they must be refueled, often multiple times. We solve the AFRP and its associated aircrew scheduling problem (Combs 2002; Combs and Moore 2004) for the resulting aircraft schedule. With the strategic mission complete, we focus on the distribution of commodities within a theater of war (Crino 2002; Crino et al. 2004; Burks 2006; Burks et al. 2008). An additional study on determining what material should be transported by cargo ships and by cargo planes was performed by McKinzie and Barnes (2006). However, since this did not deal directly with transport operations, it is not discussed in this chapter.

6.4 The 2D Bin Packing Problem (2D-BPP)

The US Armed forces have automated the process for ordering and designating items for shipment. Thus, at any time, lists of items exist that must be shipped from one location to another. However, the automation that supports the loading of these items does not provide a standard strategy to obtain effective and efficient packing solutions. Current load planning is performed either by ad-hoc methods or by trial-and-error. An automated system that provides high quality packing solutions within very short implementation horizons is greatly needed. Since the tragic events of September 11, 2001, AMC has flown on average 1000 airlift missions per month (USTRANSCOM, 2005). Given the cost of each mission is nearly $1 million, approaches to reduce the number of loads in a reasonably short amount of time merit further investigation. The methodology described in this section is a first step in achieving such a system.

The 2D-BPP is concerned with $j=1, 2, \ldots, n$ rectangular items of *width* w_j and *height* h_j. An unlimited number of identical rectangular bins of width W and height H are available. The objective is to pack all items into the minimum number of bins, in such a way that no two items overlap and the item edges are parallel to those of the bins. Items may be packed in either of their two allowable orientations. It is assumed without loss of generality that all input data are positive integers and that any item will fit into a bin in at least one of its orientations. Lodi et al. (1999a) use the notation 2D|R|F (two dimensions |rotation allowed| free packing allowed) for this problem. The 2D-BPP is known to be NP-hard (Chu and Beasley, 1998).

When considering practical sized 2D-BPPs, classical exact methods are insufficient. Most recently published work has investigated heuristic and metaheuristic approaches. Lodi et al. (2002a, 2002b) present surveys on 2D packing problems and advances in 2D packing. The first paper covers models, approximation algorithms, lower bounds and exact algorithms. The second paper discusses bounds, exact methods, heuristic approaches, and metaheuristic methods for the various classes of 2D problems. Lodi et al. (1999b) present approximation algorithms for the 2D bin packing problem (BPP).

There are several popular one-pass heuristic approaches for packing 2D rectangular items. The most basic of the 2D heuristics rely on a version of the bottom left (BL) heuristic (Baker et al., 1980). Chazelle (1983) provides the first $O(n^2)$ implementations of a one-pass BL heuristic for the 2D-BPP, but offers no detail on how to keep the data structure updated. There are several different variations for selecting where an item is placed. Chazelle (1983) used the lowest possible BL stable location and broke ties by taking the leftmost, Jakobs (1996) began at the top right and then alternated movements, first moving as far as possible toward the bottom then as far as possible to the left until the item was BL stable. Liu and Teng (1999) offer a new method that allows for representation of *any* BL stable solution. Starting from the top right corner, each piece is pushed as far down as possible until it makes contact with another item and then as far to the left as possible. These BL movements are repeated until the item can no longer be shifted further to the bottom or left. A packing pattern is represented by a permutation, π, of the items. For example $\pi = (3,2,4,5,1)$ first loads item 3, then items 2, 4, 5 and 1. This is illustrated in Figure 6.2a. Figure 6.2b shows the bottom left fill (BLF) heuristic of Chazelle (1983).

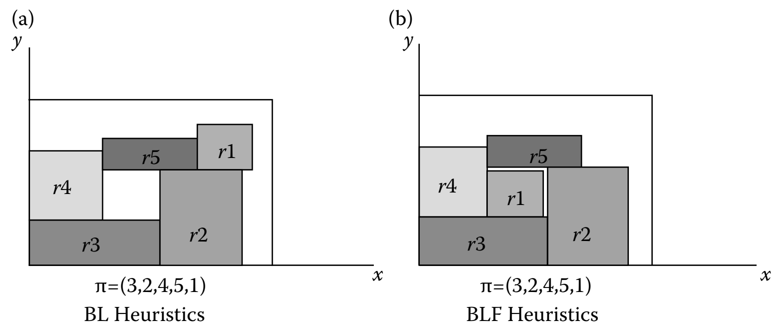


FIGURE 6.2 Examples of the BL and BLF heuristics.

In BP problems, the primary goal is to minimize the number of bins while ensuring that every item is packed. Harwig (2003) and Harwig et al. (2006) present the implementation of an advanced TS for the BP (ATS-BP) where a fine-grained objective function is developed that allows effective relative evaluation of competing moves in support of the primary goal. Moves that increase the likelihood of reducing the number of bins should be favored. Three kinds of moves fit this definition:

1. Excess bin moves. An *excess bin* is a designated candidate bin for complete emptying by moving its items to other bins. Moves that reduce the total space occupied by items in an excess bin should have high relative value. When such moves yield the same amount of occupied space, the move that results in the greater number of items in the excess bin is preferred, i.e., more items with smaller average size will be easier to relocate to other bins in subsequent rearrangements.
2. Intra-bin moves. If two moves yield different packing arrangements of the same set of items in a bin, the one with a more compact arrangement is preferred. A "compact arrangement" is discussed below.
3. Inter-bin moves. Given two bins, a move yielding greater aggregation of the dead space should be preferred because the likelihood of relocating a large item from the excess bin (or other bins) increases.

The fine-grained objective function integrates the effects of excess bin, intra-bin, and inter-bin moves. This is achieved by using a *potential energy* (PE) based approach. Li and Milenkovic (1995) used a PE based objective function for 2D strip packing. We define PE relative to *both* the x and y axes. Let μ and ρ be the vertical and horizontal gravitational constants, respectively. $CG_k=[CG_{kx},CG_{ky}]$ is the *composite CG* (center of gravity) of bin k and the items packed within bin k. CG_k is computed as

$$CG_{kx}=\frac{HW\left(\frac{1}{2}W\right)+\sum_i m_i d_{ix}}{HW+\sum_i m_i} \quad \text{and} \quad CG_{ky}=\frac{HW\left(\frac{1}{2}H\right)+\sum_i m_i d_{iy}}{HW+\sum_i m_i}$$

The PE_k of a bin k is defined analogously to that of an item, with the exception that only *open bins* (bins containing one or more items) have $PE_k>0$. The fine-grained objective function of "minimizing the total *PE* of all bins" assures that the effects of the excess bin, intra-bin, and inter-bin moves will drive the total packing configuration toward minimizing the number of packed bins. We penalize an excess bin by translating its BL corner to (2.5W, 2.5H) before determining its *PE*. (The multipliers of value 2.5 assure the excess bin's *PE* will be large relative to all other bins.) Since we do not attempt to improve the packing of an excess bin, each excess bin item is assigned a CG_i equal to the geometric center of the excess bin. However, we do assure that the items in an excess bin may be feasibly packed in that bin.

If two moves yield the same total space occupied in the excess bin, the move that leaves more items in the excess bin is preferred because smaller items are easier to place in currently open bins. This preference is enforced by subtracting $|S_i|h_{\min}w_{\min}/2HW$ from both components, where $h_{\min}$ and $w_{\min}$ are the dimensions of the smallest item. ((Lodi et al. 1999a) used a similar term in their "filling function".) The final equations that define the x- and y-values of the excess bin's *PE* are

$$PE_{\text{Excess}_x}=[3\rho W_{\text{Excess}}]\left(H_{\text{Excess}}W_{\text{Excess}}+\sum_{i\in\text{Excess}} h_i w_i-\frac{|S_i|h_{\min}w_{\min}}{2HW}\right)$$

$$PE_{\text{Excess}_y}=[3\mu H_{\text{Excess}}]\left(H_{\text{Excess}}W_{\text{Excess}}+\sum_{i\in\text{Excess}} h_i w_i-\frac{|S_i|h_{\min}w_{\min}}{2HW}\right)$$

An arrangement A is more "compact" than an arrangement B if $PE_A < PE_B$. The values selected for μ and ρ should encourage layouts that are favorable for future placements associated with intra-bin rearrangements. It is intuitively appealing to set μ = ρ. However, as illustrated in Figure 6.3, setting μ marginally greater than ρ can yield better dead space consolidation when packing in nearly full bins.

If μ = ρ, then the bin A arrangement would be preferred over the bin B arrangement ($PE_A < PE_B$). However, the bin B arrangement is superior, i.e. any item packable in bin A is packable in bin B but the *converse* is false. Consider the *PE* mathematical relationship for Figure 6.3. The *PE* relationship required to favor bin B over bin A is

$$\rho(4.5(72)+4.5(9))+\mu(4(72)+8.5(9))>\rho(4.5(72)+9.5(9)+\mu(4(72)+4.5(9)$$

which implies ρ/μ > 0.889. Bins A and B are 81% full; bins that are more full would require a higher ratio. For this reason, we use μ = 1.0 and ρ = 0.9.

Inter-bin moves should be used to consolidate dead space and the *PE* is indicative of how compactly the items are placed. However, in some cases, a simple *PE* measure can encourage inferior configurations. With two or more partially filled bins, *PE* encourages equal distribution of the items. However, a simple *PE* can favor moving an item from one bin to another which is contrary to our goal of dead space consolidation.

Modifying the *PE* measure by translating the lower left corner of a bin, X_k and Y_k, by a distance proportional to the square root of the dead space, in both the *x*- and *y*-directions, corrects this shortcoming. An additional scaling factor of 1.5 favorably weights inter-bin moves relative to intra-bin moves and ensures that the top right corner of a nonexcess bin is always at lower *PE* than the BL corner of an excess bin. Thus the lower left corners of all bins (X_j, Y_j) are redefined as

$$X_j = 1.5W_j\sqrt{1-\frac{\sum_{i\in j} h_i w_i}{H_j W_j}} \quad Y_j = 1.5H_j\sqrt{1-\frac{\sum_{i\in j} h_i w_i}{H_j W_j}}$$

After determining the new translation penalties, the new PE_k is calculated as

$$PE_k = \left[\rho(CG_{kx}+X_j)+\mu\,(CG_{ky}+Y_j)\right]\left(1+\frac{\sum_i h_i w_i}{HW}\right)(HW)$$

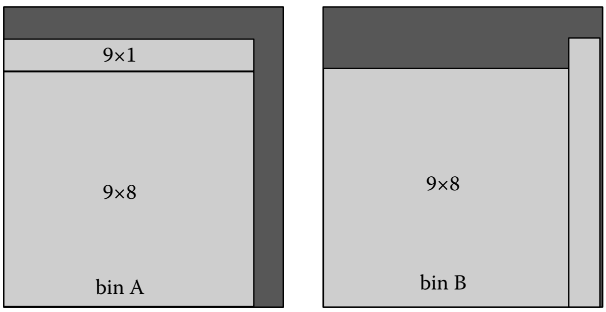


FIGURE 6.3 Adjusting μ and ρ.

Moves that reduce the contents of an excess bin also reduce the dead space penalty of the destination bin. The total *PE* of all bins comprises the fine-grained objective function.

Although this objective is tailored to the 2D BPP, the use of *PE* has potential for many types of packing problems including arbitrary shaped items in both 2D and 3D for both bin packing and geometric knapsack variants.

We use an order-based solution representation for the BP problem. Each bin is allocated a unique "letter". We reserve the numbers 0 to $n-1$ for the maximal number of bins. A bin letter, fixed as the first letter in its bin-item set, is accompanied by the letters associated with the items in the bin. Each of the n items has two associated letters. One letter represents the orientation where the item's longest axis is in the horizontal plane (orientation h). The other orientation has the longest axis in the vertical plane (orientation v). The letters n to $2n-1$ represent the orientation h for the n items in a specified order. Letters $2n$ through $3n-1$ represent orientation v for the items in the same order. One and only one orientation for each item must be in a bin-item set.

For example, suppose there are 20 items, numbered 20–39, with items 20–30 packed in order in bin 0 in orientation h excepting item 25 and items 31–39 packed in order in bin 1 in orientation h excepting item 32. The corresponding solution representation would be

(0, 20, 21, 22, 23, 24, *45*, 26, 27, 28, 29, 30) (1, 31, *52*, 33, 34, 35, 36, 37, 38, 39)

We now describe the moves used by our ATS-BP approach to the 2D-BPP. Let us consider a *single* bin and the effects of a single move on that bin. All items that are not *directly relocated* as part of the move retain their current orientation and relative order regardless of any required absolute change in the packing sequence. However, a direct result of a move may be that all items no longer fit in the bin. The pure pcf-BL manages any bin overfill by opening a new bin. The ATS-BP utilizes an excess bin for this purpose. This is achieved through a slight modification of the pcf-BL.

Whenever the ATS-BP moves an item for an insert move or a pair of items in a swap move, it *directly relocates* the selected items. However, the relocation may cause other items to no longer fit in their current bin. In this situation, the ATS-BP implements an ejection chain through the pcf-BL that feasibly places nonfitting items into an excess bin.

In terms of the solution representation, consider the following incumbent solution for the 20 item example where bin 2 is the excess bin:

(0, 20, 21, 22, 23, 24, **45**, 26, 27, 28, 29, 30) (1, 31, **52**, 33, 34, 35, 36, 37) (2, 38, 39)

Swapping items 45 and 52 initially yields

(0, 20, 21, 22, 23, 24, **52**, 26, 27, 28, 29, 30) (1, 31, **45**, 33, 34, 35, 36, 37) (2, 38, 39)

However, item 37 no longer fits in bin 1 and is ejected to the excess bin 2 yielding

(0, 20, 21, 22, 23, 24, **52**, 26, 27, 28, 29, 30) (1, 31, **45**, 33, 34, 35, 36) (2, 37, 38, 39)

In general, larger items are harder to place. At each iteration, ATS-BP identifies the excess bin's *Big-Item*, i.e. the item with maximal $h_i w_i$. Excess bin moves seek to relocate the Big-Item and other relatively large items into packing positions by checking both orientations. No items smaller than a stated threshold (relative to the Big-Item) are moved. The excess bin threshold function is detailed below.

An *excess-bin-insert move* extracts a single item from the excess bin and places it in a packing position. This move is the primary move for emptying an excess bin. At each iteration, the ATS-BP attempts to insert large items into all possible packing positions.

Excess-bin-swap moves swap *only* the Big-Item with candidate items in packing positions and may cause ejections. (Both Big-Item orientations are considered.)

Candidates to be swapped with the Big-Item are screened based on specified lower and upper values proportional to the Big-Item. Big-Items are not allowed to swap with items of identical size. These screening methods reduce the neighborhood size without sacrificing effectiveness. Excess-bin-swap moves consider only the Big-Item to enhance its likelihood of placement.

Intra-bin insert moves relocate a single item from one packing position to another in the *same* bin. Since intra-bin insert moves make small or null changes in the fine-grained objective function and are prone to cycling, ATS-BP prohibits consecutive intra-bin-insert moves. Since the pcf-BL heuristic implements a placement in the first BL stable packing position encountered, moves that change the order of items to be placed can result in a new placement in a highly desired packing position. A special case of the intra-bin-insert move, the *trapped-hole-delay* move, attempts to make such packing positions available.

A trapped-hole-delay move directly relocates an item that completely "seals" an unoccupied space, to a later position in the packing order. Trapped-hole-delay moves are less prone to cycling and are allowed to be performed up to two consecutive times. Only bins with more than five items are considered for this move.

The inter-bin moves, inter-bin-insert or inter-bin-swap, allow ATS-BP to consolidate dead space by moving at least one item from a position in one bin and placing it into another. The *inter-bin-insert move* inserts one item, considering both orientations, from one packing position to a packing position in another bin. The departure bin is repacked without the relocated item. The heuristic attempts to repack the destination bin with the inserted item. The move is not considered if the inserted item ejects from the destination bin. However, items that follow the inserted item are allowed to eject. This move only considers items that either have $h_i<h_{\text{Big}}$ or $w_i<w_{\text{Big}}$. This is the most powerful of the nonexcess bin moves since it can empty an open bin.

The *inter-bin-swap move* swaps the packing positions of two items in different bins. Ejection of either of the moved items is not allowed. Many items too large for inter-bin-insert moves are accessible to inter-bin-swap moves. Identical items are not considered for swap moves. Inter-bin-swap moves embody a search intensification context which preserves the bin orientation of the swapped objects.

ATS-BP imposes two types of tabu restrictions, tabu bin and tabu order.

Tabu bin prohibits an item, in either orientation, from returning to an excess bin within tabu tenure iterations. The tabu bin attributes consist of a bin letter and an item's letters. *Tabu order* prohibits an item-orientation from repeating a particular packing order associated with a bin, i.e. prohibits an item-orientation letter from returning to a specific position in a bin within tabu tenure iterations. Tabu order is used for all moves not involving excess bins. A triplet of tabu attributes is used for this restriction, (bin letter, item letter, item position). Moves satisfying an *aspiration criterion* can override a tabu restriction. ATS-BP uses a single aspiration criterion; only moves that achieve a new best global objective function value are allowed to override a tabu restriction.

The tabu bin and tabu order restrictions use the same *adaptive tabu tenure structure*. The initial tenure is set to a minimum value. If the selected move does not improve the objective function, the tabu tenure increases by 1. If the selected move improves the objective function, the tabu tenure decreases by one. The increase (decrease) is never allowed to violate a maximum (minimum) tenure. Any time an aspiration criterion is met, the tabu tenure is reset to the minimum.

Before packing any bins, a simple lower bound on the number of bins, *SLB*, and an *item-to-bin ratio*, *IBR*, are determined. Define

$$SLB = \frac{\sum_{i}^{n} h_i w_i}{HW}$$

where n is the total number of items, h_i is the height and w_i is the width of item i. H and W are the common height and width of the set of identical bins. The item to bin ratio, $\text{IBR}=n/SLB$, is an upper bound on the average number of items that can be packed in a bin and is used for bin selection.

ATS-BP has two minimum and maximum tenure settings associated with the IBR:

1. IBR≤7: minimum seven and maximum 12.
2. IBR>7: minimum nine and maximum 14.

These limits reflect the fact that the higher IBR problems make greater use of cycle prone intra-bin moves and higher IBR problems have greater likelihood to generate identical geometric arrangements with the different BL solutions.

We now discuss the dynamic move selection strategies and the logical flow of the heuristic. Key to implementing a successful move strategy is defining high and low influence moves. Influence "… measures the degree of change induced in solution structure..." (Glover and Laguna, 1997) High influence moves in ATS-BP are the excess bin moves. All other moves have low influence. If a nonejecting excess-bin-insert move is found or if any excess bin move finds a new global best solution relative to the fine-grained objective, the best high influence move is implemented. ATS-BP allows a maximum of four consecutive low influence moves. ATS-BP rewards low influence moves that yield arrangements favorable to future high influence moves. It also limits, independently, both the number of consecutive inter-bin moves and intra-bin moves.

The dynamic move selection strategies vary the frequency of use and the size of both high and low influence move neighborhoods in order to achieve intensification/diversification effects. ATS-BP selects which move neighborhoods are implemented and changes the neighborhood sizes based on the following measures:

1. Number of items, *n*.
2. IBR.
3. Aspiration riteria count (AC): number of iterations since the global best solution was found.
4. Nonimproving moves count (NIMC): number of iterations since last improving move.
5. Nonexcess bin moves count (NEBMC): number of consecutive low influence moves.
6. Inter-bin moves count (InterMC): number of consecutive inter-bin swaps or inserts.
7. Intra-bin moves count (IntraMC): number of consecutive hole-top inserts or intra-bin inserts.
8. For problems with IBR < 7, the search strategy oscillates (every 50 iterations) between the intensification and diversification modes, starting with the intensification mode. The ATS-BP uses *only* the intensification mode for problems with IBR ≥ 7. These modes are discussed below.

We observed that the number of bins relative to the number of items has a noticeable affect on which move types solve the problem more quickly. For example, for low IBR, the inter-bin moves are quite successful at traversing the landscape. However, as the bin contains more move items, the intra-bin moves become essential. However, the intra-bin move must be regulated because of the enhanced potential for cycling due to the prevalence of identical geometric arrangements.

All excess-bin moves are evaluated at each iteration. The *excess-bin-insert neighborhood* incorporates a dynamic *functional* candidate list mechanism to screen items considered for extraction from the excess bin, i.e. only items that satisfy

$$(h_i > \varphi\, h_{\text{Big}} \cup w_i > \varphi\, w_{\text{Big}}) \cap (h_i w_i \geq f(\varphi) h_{\text{Big}} w_{\text{Big}})$$

are considered, where $f(\varphi)$ is a discrete function of φ as detailed below.

The Big-Item is *always* a candidate for excess-bin extraction. The logic supporting this pair of relations is that a high value of φ will strongly restrict the candidate list of items available for insertion from the excess bin and will favor larger items *dissimilar* from the Big-Item. Small values of φ will have the opposite effect, admitting more items to this candidate list while still prohibiting the smallest items. $\varphi \equiv 1$ is the default setting for ordinary search conditions. $\varphi \equiv 2$ is applied when the number of items in the excess-bin exceeds 12. This prevents the larger items from being overshadowed by numerous smaller items and hindering the progress of the search, while still allowing other large items with notably different shapes to be move candidates. $\varphi \equiv 0.8$ and $\varphi \equiv 0.6$ are increasingly liberal values which allow larger candidate lists. φ is set to 0.8 to combat moderate search stagnation evidenced by three simultaneous conditions: no recent aspiration criterion satisfaction, a nonimproving previous move, and a low influence previous move. φ is set to 0.6 when a larger number of iterations have seen no aspiration criterion satisfaction and the last two moves were low influence moves.

The discrete function, $f(\varphi)$, was created for the following reasons. When the search is not in the diversification mode and $\varphi \neq 2$, $f(\varphi) = 0.5\varphi$, which achieves the standard area threshold for candidate list membership. When the search is in the diversification mode and $\varphi \neq 2$, $f(\varphi)$ is set to 0.65φ to be more restrictive on small items becoming part of the candidate list for excess-bin-insert moves. When $\varphi = 2$, $f(\varphi)$ is set to 0.375φ to relax the restriction on area, since $f(\varphi) = 0.5\varphi$ would force candidate items to be of identical area.

The *excess-bin-swap neighborhood* uses a single static equation to dynamically screen candidates for swapping positions with the Big-Item in the excess bin,

$$\frac{1}{16} h_{\text{Big}} w_{\text{Big}} \leq h_i w_i \leq 4 h_{\text{Big}} w_{\text{Big}}$$

where all items, i, in any packing position are considered. The screening method is dynamic because the Big-Item may change with each iteration. Since this neighborhood is evaluated at every iteration, this restriction is present to lessen the associated computational effort without lessening the power of the search. If the best move from this neighborhood achieves the aspiration criterion, the remaining neighborhoods are not checked.

Having considered the high influence excess bin moves, the algorithm proceeds to the low influence moves. If the last four iterations have used low influence moves, the best high influence move is immediately implemented. This is to reduce the chances of cycling and force a change in the candidate moves in the excess bin. Otherwise, if the last two moves have *not* been intra-bin moves (which include intra-bin-insert and trapped-hole-delay moves) and $\text{IBR} \geq 5$, the trapped-hole-delay neighborhood is evaluated. In this neighborhood, bins with more than five items are checked for trapped holes and the move is conducted on all items which seal trapped-holes. The IBR threshold prevents this type of move computation when associated improvements are unlikely.

Intra-bin-insert moves are examined only if the previous move was not an intra-bin move. In addition, when the search is in the diversification mode (which only occurs when $\text{IBR} < 7$), the additional restriction of the last three moves not being low influence moves is imposed. In evaluating the intra-bin neighborhood, candidate bins are limited to those bins with a specified amount of dead space in any candidate bin.

We now discuss the inter-bin neighborhoods. If the InterMC, which includes both inter-bin-insert and inter-bin-swap moves, is equal to three, the inter-bin neighborhoods are not evaluated. This criterion controls an over application of inter-bin moves while allowing essential consolidation of dead space. Given that $\text{InterMC} < 3$, one addition condition can preclude evaluation of the inter-bin neighborhoods:

$$\text{diversification mode active} \cap n \leq 60 \cap \text{InterMC} = 2$$

This additional restriction is present because the smaller problems more quickly achieve over application of this move type.

When the inter-bin neighborhoods are evaluated, inter-bin-insert moves are considered first. Inter-bin-insert moves consolidate dead space possibly creating adequate space for relocation of the Big-Item in the excess bin. When NEBMC<3, candidate moves are required to create dead space in the departure bin that is at least 50% of the area of the Big-Item. When NEBMC=3, candidate moves are required to create dead space in the departure bin that are greater than or equal to the area of the Big-Item. When AC>10 and NEBMC<2, the dead space requirement is no longer enforced with the goal of enabling a short term diversification. If the inter-bin-insert neighborhood finds a new global best solution, the related move is implemented without further investigation. Failing a new global best solution, the algorithm continues with the evaluation of the inter-bin-swap neighborhood.

The number of candidate moves for the inter-bin-swap neighborhood depends on NEBMC, AC, and whether the search is in the diversification mode. The *combined dead space* of candidate bin pairs (j,k), must exceed a defined proportion of the Big-Item area. During intensification and when the NEBMC≥1, candidate moves must create a dead space area in one of the bins exceeding the size of the Big-Item. For problems with high IBR, this is easily achieved. For lower IBR problems, this equation greatly reduces the candidates that must be examined.

If the search is in the diversification mode and NEBMC=0:

1. The maximum dead space constraint is not applied
2. Candidate moves are required to have combined dead space that is at least 50% of the area of the Big-Item.
3. When AC>10, the combined dead space requirement is no longer enforced with the goal of enabling a short term diversification.

After consideration of all possible moves, the "best move" is implemented, the tabu structures are updated, and the tabu tenure and all counters are adjusted. If required, a new excess bin and/or Big-Item is determined. The ATS-BP program continues iterations until a stopping criterion is reached.

One final comment about the ATS-BP structure involves an efficient bin data structure that appears to have been overlooked in previous research. The pcf-BL heuristic is an efficient 2D packing heuristic requiring $O(m^2)$ operations for each move evaluated, where m is the number of items that are in bins to be repacked. Since the low influence moves have up to $O(n^2)$ combinations, the number of low influence moves per iteration grows rapidly with problem size. Techniques to reduce the number of moves considered will enhance computational efficiency. In addition to the basic screening criteria detailed in the previous section, this research implements two novel approaches to improve efficiency that can be adapted to any direct search method using an order-based heuristic applied to multidimensional bin packing.

Dowsland (1996) states that order based heuristics require a complete repacking of the container at each move. This is true only in a very rare worst case scenario. The ATS-BP saves the bin data structure as each item is packed so that future moves do not have to completely repack a bin. The bins are only repacked from the point of change. This is achieved by storing the "Bin History" for each item as it was packed. (A numerical Item Data list is also stored that saves where each item is packed.) The Bin History saves all crucial arrays and data needed to continue the pcf-BL from any point in the packing sequence. Whenever a new move is accepted, the bin history for all accepted items is updated. Whenever an item is inserted into a sequence, the ATS-BP identifies its immediate predecessor and retrieves the Bin History for the item. The expected computational savings associated with this technique is 50%. Throughout the above discussion, we have attempted to explain the rational supporting the various components in the ATS-BP. This is especially true in the case of the set of constant parameters that are used at several points of the algorithm construction. The complexity of the 2D orthogonal packing problem demands an algorithm that is able to embrace that complexity. For this reason, the ATS-BP uses a set of constant parameters whose values have been selected from a strategic view bolstered by our in-depth understanding of the problem at hand.

We now present the ATS-BP results for the Berkey and Wang (1987) and Martello and Vigo (1998) test sets, compare those results with those of Lodi et al. (1999a). The test sets comprise 10 different BPP classes. The first six classes are from the Berkey and Wang (1987) test set. The last four are those proposed by Martello and Vigo (1998). Each class considers five values of n: 20, 40, 60, 80, and 100. Each n has ten instances for a total of 500 problems.

The six classes of the Berkey and Wang (1987) test set are described below:

Class I: w_j and h_j uniformly random in [1,10], $W=H=10$.
Class II: w_j and h_j uniformly random in [1,10], $W=H=30$.
Class III: w_j and h_j uniformly random in [1,35], $W=H=40$.
Class IV: w_j and h_j uniformly random in [1,35], $W=H=100$.
Class V: w_j and h_j uniformly random in [1,100], $W=H=100$.
Class VI: w_j and h_j uniformly random in [1,100], $W=H=300$.

The Martello and Vigo test set (Lodi et al. 1999a) was constructed to consider "… a more realistic situation ..." where not all items are generated on the same interval. Four *types* of items are combined in varying proportions as described below:

Type 1: w_j uniformly random in [2/3W, W] and h_j uniformly random in [1,1/2H].
Type 2: w_j uniformly random in [1,1/2W] and h_j uniformly random in [2/3H,H].
Type 3: w_j uniformly random in [1/2W, W] and h_j uniformly random in [1/2H,H].
Type 4: w_j uniformly random in [1,1/2W] and h_j uniformly random in [1,1/2H].

These item types created the following four problem classes:

Class VII: *Type 1* with probability 70%, *Type 2, 3, 4* with probability 10% each.
Class VIII: *Type 2* with probability 70%, *Type 1, 3, 4* with probability 10% each.
Class IX: *Type 3* with probability 70%, *Type 1, 2, 4* with probability 10% each.
Class X: *Type 4* with probability 70%, *Type 1, 2, 3* with probability 10% each.

ATS-BP results detailing the number of bins used and solution times are provided in Harwig (2003) for *each* of the 500 instances. Unfortunately, the *only* comparable results available for the previous best technique are the average ratios for each 10 problem subset as reported for Lodi et al.'s (1999a) TS-TP method.

Tables 6.1 and 6.2 detail the comparative average results for the ATS-BP and for Lodi et al.'s (1999a) TS-TP method. The results in Tables 6.1 and 6.2 are given in terms of an *average* ratio across the 10 problem instances in each subclass where the numerator is the number of bins used by the heuristic and the denominator is the improved lower bound (LB) on the required number of bins from Dell'Amico et al. (2002). The consolidated results (where all problems are weighted equally) are presented for each test set in the row labeled "Test set average" in Tables 6.1 and 6.2. Overall, the ATS-BP achieved an average ratio of 1.045. This improvement of 0.015, compared to the TS-TP result of 1.060 reported by Lodi et al. (1999a), is a *25% reduction* in the previous best average result.

The largest improvements occurred in the three test sets with the highest IBR. Classes II, IV, and VI had IBR's that ranged from a minimum of 20 to a maximum of 40. For these three classes, the ATS-BP found results within 2.1% of the LB, while TS-TP was within 5.7% of the LB. Part of this difference is explained by the TS-TP's inability to improve an initial solution to a single bin problem, and its limited search neighborhood when there are only a few bins.

ATS-BP also performed well on data sets with low IBR. ATP-BP was within 5.5% of the LB while TP-TS was within 6.1% of the LB. Classes I, III, V, VII, VIII, IX, X have IBR values ranging from 2 to 10. However, of those classes only Class X had problems where the IBR exceeded 5.

ATS-BP did take longer to find the answer to some of the problems that TS-TP discovered rather quickly. This is not surprising. Methods that sort by size tend to perform well in many cases. If a "sorted" good solution exists, these methods quickly find it. In general, there are $m!2^m$ orderings for a set of m items to be ordered for packing in a single bin. Of these, TS-TP considers only one ordering. However, in many instances that one ordering is all that is required.

TABLE 6.1 Berkey and Wang Test Set Consolidated Results

Berkey and Wang Test Set (1987) (2BP|R|F)						
Lodi, Martello, and Vigo (1999)						
		TP TS		ATS-BP		
Class	*n*	LB	Average Time	LB	Average Time	Time to Match
I	20	1.05	18.00	1.03	0.13	0.13
	40	1.04	30.02	1.04	4.11	4.11
	60	1.04	34.00	1.04	4.27	4.27
	80	1.06	48.08	1.06	0.97	0.97
	100	1.03	47.92	1.03	1.20	1.20
	Average	1.044	35.60	1.040	2.14	2.14
II	20	1.00	0.01	1.00	0.13	0.13
	40	1.10	0.01	1.00	1.46	0.21
	60	1.00	0.01	1.00	1.14	1.14
	80	1.03	6.10	1.00	22.61	2.76
	100	1.00	0.01	1.00	4.25	4.25
	Average	1.026	1.23	1.000	5.92	1.70
III	20	1.06	18.00	1.02	11.35	0.17
	40	1.09	42.17	1.09	0.38	0.38
	60	1.08	54.15	1.08	64.85	64.85
	80	1.07	60.07	1.07	9.94	9.94
	100	1.07	60.18	1.06	68.05	0.98
	Average	1.074	46.91	1.064	30.91	15.26
IV	20	1.00	0.01	1.00	0.14	0.14
	40	1.00	0.01	1.00	0.25	0.25
	60	1.10	0.09	1.05	28.80	0.31
	80	1.07	12.00	1.07	36.16	36.17
	100	1.03	6.00	1.03	9.19	9.19
	Average	1.040	3.62	1.030	14.91	9.21
V	20	1.04	12.01	1.04	1.31	1.31
	40	1.07	42.00	1.06	41.21	18.72
	60	1.06	45.23	1.06	74.59	74.59
	80	1.07	54.14	1.06	143.01	14.63
	100	1.07	60.12	1.05	196.23	1.58
	Average	1.062	42.70	1.054	91.27	22.17
VI	20	1.00	0.01	1.00	0.17	0.17
	40	1.40	0.03	1.10	5.97	0.27
	60	1.05	0.05	1.00	25.92	0.32
	80	1.00	0.01	1.00	0.42	0.42
	100	1.07	12.00	1.07	7.64	7.64
	Average	1.104	2.42	1.034	8.02	1.76
Test set average		1.058	22.081	1.037	25.528	8.707

6.5 3D Pallet Packing

This problem involves packing a rectangular pallet of given length, width and height dimensions with a set of rectangular boxes. The set of rectangular boxes contain various box types and all boxes of a particular type have the same dimensions. Each box type is defined by four parameters: length, width, height and the number of boxes of that type. Our goal is to pack selected boxes onto the pallet, within

TABLE 6.2 Martello and Vigo Test Set Consolidated Results

Martello and Vigo Test Set (1998) (2BP|R|F)						
Lodi, Martello, and Vigo (1999)						
		TP TS		ATS-BP		
Class	*n*	LB	Average Time	LB	Average Time	Time to Match
VII	20	1.11	30.00	1.11	0.77	0.77
	40	1.08	48.06	1.07	99.28	36.93
	60	1.06	59.45	1.04	613.98	115.98
	80	1.10	60.12	1.08	259.78	54.06
	100	1.08	60.36	1.07	231.66	30.08
	Average	1.086	51.60	1.074	241.09	47.56
VIII	20	1.10	30.01	1.10	1.34	1.34
	40	1.10	54.22	1.08	77.91	0.28
	60	1.07	56.17	1.07	51.74	51.74
	80	1.08	60.11	1.07	529.40	10.69
	100	1.09	60.14	1.08	232.88	0.72
	Average	1.088	52.13	1.080	178.65	12.95
IX	20	1.00	0.06	1.00	0.26	0.26
	40	1.01	18.85	1.01	1.25	1.25
	60	1.01	18.03	1.01	2.54	2.54
	80	1.01	30.51	1.01	1.65	0.65
	100	1.01	36.86	1.01	52.69	1.39
	Average	1.008	20.86	1.008	11.68	1.22
X	20	1.12	6.01	1.12	0.36	0.36
	40	1.06	24.01	1.06	3.93	3.88
	60	1.06	30.44	1.06	235.17	235.17
	80	1.05	39.04	1.05	78.27	78.27
	100	1.05	43.38	1.04	269.68	12.25
	Average	1.068	28.58	1.066	117.48	65.99
Test set average		1.063	38.292	1.057	137.227	31.931
Grand average		1.060	28.565	1.045	70.208	17.996

the pallet volume to minimize the unused space in the pallet volume. We impose logical constraints to ensure no overlapping of boxes and packing within the defined dimension limits of the pallet. Box rotation is allowed and boxes are packed using any of six orthogonal orientations. We do not currently consider box weight or pallet center of gravity (CG); the current focus is pallet volume utilization. The United States Air Force (USAF) uses standard HCU-6/E (463L) pallets in air shipping. The length and the width of the pallets are 88″ and 108″, respectively. The maximum height of a pallet is 96″ if the pallet is loaded in an aircraft's main compartment and 76″ if the pallet is loaded on the aircraft's loading ramp (Ballew, 2000). Our packing algorithm presented in Baltacioğlu (2001) and Baltacioğlu et al. (2006) considers pallet packing in any orientation, both in terms of the pallet and the boxes. The reader should note this algorithm is not a TS approach, but it is a heuristic. We treat orthogonal pallet orientation (of which there are six) as an outer processing loop, while each viable packing layer thickness of a packing orientation is treated as an inner processing loop. Thus, within every iteration of our packing algorithm, each orientation of the pallet is considered when determining packing layer specifications and particular component object placement. We examine the set of candidate objects and determine a most viable "thickness" to start the packing of each layer of objects in the restricted space. We then choose candidate objects according to that packing layer. Our packing approach includes a sub-algorithm that defines each of the potential layers and then, starting at the smallest (and most favorable) layer level, considers each layer level in turn to start each of the packing iterations. Our general approach is described by the following pseudo code:

```
For each pallet orientation
For each layer thickness
Pack boxes
Next
Save packing obtained
Next
Return best packing obtained
```

The 3D packing problem (de Castro Silva et al. 2003; Lodi et al. 2002b; Martello et al. 2000; Szykman and Cagan 1997) is a generalization of the 1- and 2D packing problems and is NP-hard (Reeves 1995). Unlike the container loading problem (Faina 2000), our problem's container or pallet has a fixed height. In general, optimal solutions are computationally impractical to achieve, so several heuristics have been developed (Bischoff et al. 1995; Gehring and Bortfeld 1996; Gehring et al. 1990; George and Robinson 1980; Han et al. 1989; Lim et al. 2003; Liu and Chen 1981; Loh and Nee 1992; Pisinger 2002; Terno et al. 2000; Wu et al. 2002). Most studies have focused on the practical aspects of loading a container or pallet and developing heuristic solutions based on the concept of filling out the container or pallet with boxes organized in layers, walls and columns (Bischoff and Dowsland 1982; Bischoff and Marriott 1990; George and Robinson 1980; Lim et al. 2003; Liu and Chen 1981). These heuristics are, in general, online packing algorithms, which means they pack boxes one-by-one in a given order. More precisely, when the algorithm is packing a box, it has information only about the boxes previously packed, and once a box is packed, that box cannot be moved to another place on the pallet. Heuristics and algorithms require efficient data structures for efficient execution. Our data structures consist of two arrays and a double linked list. The Boxlist[] array contains 12 fields per record, one record per box. Each record stores all box dimensions, the coordinates where it is packed on the pallet, as well as the volume of the box. Additional fields can easily be added, if necessary, to facilitate particular implementations of our algorithm.

The Layers[] array is created for each pallet orientation and contains every unique dimension of the boxes less than the y dimension of the current orientation of the pallet along with an individual evaluation value. Each unique value in this array represents a layer thickness value with which each iteration can start a packing. The evaluation value associated with each layer thickness value represents how well the candidate boxes match this layer thickness when this layer thickness value is used in the packing. The calculations used are a novel approach so we next provide a detailed example.

Let y^* be the y value of the current pallet orientation, B the set of boxes considered, $b \in B$ any box. We use $b_i \in B$ when referring to a single box. Denote the x, y, and z dimension attributes of each $b \in B$ by $b(x)$, $b(y)$, and $b(z)$, respectively. Define L, W, and H as sets of x, y, and z box dimensions, respectively, for $b \in B$. Clearly, $b(x) \in L$, $b(y) \in W$, and $b(z) \in H$. Define the *PLayers* set as $\{a \in L \cap W \cap H \mid a \leq y^*\}$ and $J = |PLayers|$ the size of *PLayers*.

Let Layers[] have J rows and two columns. The first column contains each of the J elements of *PLayers*, the layer thickness values. The second column contains evaluations of each corresponding layer thickness value defined as:

$$\text{Layers}[\bullet, 2] = \left\{ \sum_{b \in B} \alpha_b \mid y_o = \text{Layers}[\bullet, 1],\ \alpha_b = \min\left(|b(x) - y_o|,\ |b(y) - y_o|,\ |b(z) - y_o|\right) \right\}$$

The evaluation values measure the "smoothness" of a layer thickness value if that thickness value is used as a packing layer thickness value.

Given four boxes with dimensions 70, 104, and 24; two boxes with dimensions 14, 104, 28; three boxes with dimensions 40, 52, 36; and a pallet orientation of $x=104$, $y=96$, and $z=84$, $y^*=96$, *PLayers*={14, 24, 36, 40, 48, 52, 70}. Consider Layers[1, 1]=14 and compute Layers[1, 2] as

$$|24-14|+|24-14|+|24-14|+|24-14|+|\;|+|14-14|+|14-14|+|36-14|+|36-14|+|36-14|=106$$

Similar calculations for each element of *PLayers* yields the Layers[] array depicted in Figure 6.4.

Each iteration examines each layer thickness value starting with that value having the smallest evaluation value, as this may be the most suitable layer thickness value. Packing of the pallet ties closely to the six orientations of the pallet considered. Each orientation of the pallet is treated as a pallet to pack. The algorithm starts each iteration with the pallet empty, all boxes available, a pallet orientation and a layer thickness from the associated Layers[] array. During each iteration, either the starting layer thickness or the pallet orientation changes. Thus, each iteration represents a complete packing of the pallet. The algorithm saves the parameters of the best packing found, based on volume packed, to later reconstruct the packing. While the Layers[] array entries are used to start the first layer of the packing associated with each iteration, subsequent layers in that packing require calculation to determine good layer thickness values. These calculations duplicate those described for the Layers[] array entries, but apply only to the remaining candidate boxes. As before, the layer thickness with the smallest evaluation value becomes the initial layer thickness for subsequent packing layers. Layering continues until the pallet can no longer accommodate further layers, at which time the pallet is considered packed. Each packing tracks the volume of boxes packed and the volume of boxes not packed to derive a pallet utilization measure and percentage of packed box volume. Each packing also has an associated pallet orientation and Layers[] array index. The best packing found becomes the final solution returned by the algorithm. The algorithm performs layer packing or wall building as the pallet is packed, thus reducing the problem to a systematic series of 2D packing problems. The algorithm packs along the *x*- and *z*-axis keeping the *y*-axis as level as possible. To track the current topology, as each box is packed, the box's right corner coordinate is maintained in a (doubly linked) coordinate data list. As boxes are packed, this list changes. This approach means the algorithm only needs to track the current packing edge and ensures the algorithm avoids overlaps of layers and pallet edges. Thus, our algorithm focuses on a 2D problem while being cognizant of the third dimension, the *y*-axis, and using layer-in-layer to fix *y*-axis irregularities that may arise. For each packing action, the algorithm finds the smallest *z*-value in the coordinate data list and the current gap in the layer to fill. The algorithm analyzes the unpacked boxes and seeks, in precedence order, a box with a *y*-dimension closest to the current layer thickness value but not more than the maximum available *y*-dimension of the current gap to be filled, with an *x*-dimension closest to, but not more than, the maximum available *x*-dimension of the current gap to be filled and a *z*-dimension closest to the *z*-dimension of the current gap to be filled, but not more than the maximum available *z*-dimension of the current gap to be filled. This means it considers the *y*-dimension, then among boxes having the same *y*-dimension, it looks at the *x*-dimension, and finally among the boxes having the same *y*- and *x*-dimension, it looks at the *z*-dimension. The algorithm calculates the differences between the gap's dimensions and the box dimensions for each box and packs the box with the smallest difference.

	Layers[*,1]	Layers[*,2]
1	14	106
2	24	56
3	36	72
4	40	80
5	48	100
6	52	80
7	70	98

FIGURE 6.4 Creating the layers [] array.

The algorithm may build an uneven layer (uneven in the y-axis). Thus, our use of layer-in-layer packing better utilizes the resultant gaps. The algorithm involves layer-in-layer if it determines that there is unevenness in the third dimension of the current packing layer. To accomplish the layer-in-layer packing, the current layer thickness is increased to the largest packed y-dimension. When the current layer thickness is increased, the total increment of the layer, from the beginning to the end of the current layer packing, is stored. After completing the packing of the current layer, another packing within the layer is performed using the stored value as layer thickness. We employ various techniques to empirically test our algorithm. Bischoff et al. (1995) generated 700 problems that are available from the Imperial College Management School web page (Beasley 2005). These problems are randomly generated and come in seven files. Each file contains 100 problems with 3, 5, 8, 10, 12, 15, and 20 box types, respectively. All problems use the pallet dimensions of 587, 233, and 220.

Table 6.3 presents our solutions along with the solutions of Bischoff and Ratcliff (1995) (denoted by B/R); Gehring and Bortfeld (1996) (denoted by G/B); Bortfeld and Gehring (1997) (denoted by B/G); and Terno et al. (2000) (denoted by T/S/S/R). Information about the G/B and the B/G results come from Terno et al. (2000).

Our average and maximum utilization results compare well with the best among the four others. Note all of the minimum values produced by our algorithm are 1–4% better than those of the other approaches. Solution time is another important factor. Our algorithm solves any of 700 box sets in 1–12 sec on a Pentium III 750 MHz computer. Time comparisons with the other works are unfortunately unavailable.

Overall, our algorithm again did quite well, both in terms of processing time and pallet volume utilization. Our algorithm is particularly novel as it creates a packing in layers, fills gaps in any layer, and examines various box orientations while packing within a given volume. We implemented and rigorously tested our algorithm to include using visualization techniques to verify our approach, using attributive memory and dynamic data structures to improve processing efficiency. Our packing algorithm is both novel and highly competitive with existing methods despite our current exclusion of a local improvement (LI) phase.

6.6 Aircraft Loading

We now deal with partitioning the pallets into aircraft loads, selecting an efficient and effective subset of aircraft from an available pool of aircraft to carry the aircraft loads, and feasibly placing the pallets in the best allowable positions on the aircraft. These tasks are interdependent and their combination augmented with the temporal considerations of when cargo is available for loading at its APOE and when it should be delivered to its designated APOD, composes what is defined here as the airlift loading problem (ALP). Our approach assumes the pallets are immediately available for loading, so we model the static ALP (SALP). The SALP differs from many partitioning and packing problems described in the literature because, in addition to spatial constraints, factors such as Allowable Cabin Load (ACL) and balance restrictions must be considered. ACL refers to the maximum payload that can be carried on some aircraft mission, and balance restrictions require that the center of balance (CB) of the loaded aircraft must be within specified tolerances to ensure safe flight. Detailed in Roesener (2006) and Roesener et al. (2008), our model, SALP-TS, solves the entire problem, in a timely manner, by employing an advanced TS approach.

Research conducted on the ALP, largely limited to the military and to companies conducting research for military agencies, has not been extensive. Researchers have approached the ALP from different perspectives rendering meaningful algorithmic comparisons difficult. A short synopsis of relevant research is provided here. Cochard and Yost (1985) developed the Deployable Mobility Execution System (DMES) for use by the USAF. Their model used a modified cutting stock heuristic which only generated feasible loads for the aircraft. These feasible loads could then be modified by the load planner using an interactive user interface. Computer Aided Load Manifesting (CALM), an augmented version

TABLE 6.3 Comparisons for the Bischoff and Ratcliff (1995) Examples

	B/R			G/B			B/G			T/S/S/R			OURS		
Set #	Minimum	Average	Maximum	Minimum	Average	Maximum	Minimum	Average	Maximum	Minimum	Average	Maximum	Minimum	Average	Maximum
BR-3	73.7	85.4	94.4	76.7	85.8	94.3	78.7	89.0	95.7	75.7	89.9	95.9	78.9	89.0	95.3
BR-5	73.8	86.3	93.8	78.4	87.3	95.2	79.7	88.7	95.0	81.9	89.6	94.7	84.8	89.0	94.0
BR-8	75.3	85.9	92.6	81.1	88.1	92.9	82.4	88.2	94.0	83.2	89.2	93.0	84.5	88.4	92.1
BR-10	78.4	85.1	90.1	82.7	88.0	91.6	80.9	87.4	92.0	83.1	88.9	92.7	84.4	88.2	91.9
BR-12	78.7	85.2	90.4	81.7	87.9	92.6	81.2	86.7	91.4	83.0	88.3	91.6	84.0	87.6	89.9
BR-15	75.2	83.8	89.2	84.1	87.9	92.5	81.4	85.5	90.4	82.3	87.4	90.5	84.3	87.4	91.3
BR-20	75.7	83.0	88.3	84.4	87.7	90.7	79.9	83.9	88.4	80.6	86.3	89.0	84.3	87.1	90.2

of DMES, became the USAF standard but was later replaced due to its inadequacies in the generation of solutions to large scale SALPs. In 1998, the DOD mandated that all US military branches use the Automated Air Load Planning System (AALPS) when planning all airlift missions. AALPS is a very quick and user friendly "knowledge-based expert system" that assists users in planning and executing aircraft loads for all types of deployments (Automated Airload Planning System, http://www.tis.army.mil/AALPS/default.htm). AALPS ensures feasible aircraft loads and produces an upper bound on the number of aircraft needed to transport a set of cargo items. Unfortunately AALPS utilizes a loading schema which will often use more aircraft than are actually needed. To accomplish its loading algorithm, AALPS first sorts the cargo items by decreasing priority and then by decreasing weight. If all items have the same priority, they are simply sorted by decreasing weight. AALPS iteratively loads the heaviest item with the highest priority. When the cargo load in the aircraft reaches either the aircraft ACL or maximizes the available space, the algorithm continues on to loading the next aircraft. Chocolaad (1998) used TS to approach the SALP by assigning cargo utilities by decreasing weight. Chocolaad hierarchically solved two subproblems: a knapsack problem (KP) and a packing problem (PP). The KP selected the items and the PP placed them. Note that heavier objects were deemed more important than lighter objects, which is not necessarily the case in real-world applications. Chocolaad considered the longitudinal axis CB of the aircraft but ignored the vertical and lateral axis CB because changes in these axes "are small and flight controls can compensate for any effect on the stability as unimportant." Chocolaad's algorithm was applicable only to the C-17A Globemaster III (C-17) and considered only one aircraft at a time. Romaine (1999) extended Chocolaad's work to also use the C-5B Galaxy (C-5) and allowed the loading of up to seven aircraft. Modeling the SALP as a 2D bin packing problem (2D-BPP) (Garey and Johnson, 1979; Lodi et al. 2002a, 2002b; Eglese, 1990; Raidl and Kodydek, 1998; Bhatia and Basu, 2004; Harwig et al. 2006) is an attractive but naïve concept. Heidelberg et al. (1998) is representative of research that has approached the SALP as a 2D-BPP (ignoring height constraints) using length and width of the cargo items and of the aircraft's cargo hold. They indicate that classical methods of bin-packing are inadequate in aircraft loading because they ignore aircraft CB concerns and the fact that pallets cannot be positioned immediately adjacent to another pallet because they require tie-down restraints to ensure their stability in flight. Guéret et al. (2003) approached the SALP as a bi-dimensional BPP, with several bins and several additional constraints. In their two-phased method, the first phase consists of two possible heuristics that quickly compute "good" initial solutions; the second phase is a local search algorithm that attempts to improve upon the initial solution. Their algorithm ignores the aircraft CB.

Given a set of pallets to be transported and a set of available aircraft, the goal of the SALP is to transport all of the pallets using the fewest number of aircraft possible while satisfying the space and weight restrictions and minimizing the CB deviations of each aircraft. The SALP presumes all cargo items are palletized and have the same destination. In addition, temporal constraints are ignored and each aircraft is allowed a single use (trip).

The Department of Defense (DOD) uses standard 463L pallets for transportation of bulk cargo. The 463L pallet, of dimensions 88″×108″×2¼″, has a wood core and is covered with corrosion-resistant aluminum. It has 22 tie-down rings attached for cargo netting which requires 2″ on all pallet sides. The 463L pallet has a maximum weight capacity of 10,000 lbs with no more than 250 pounds per square inch allowed. Practical considerations require that pallets be loaded to at least 2500 lbs. Each loaded pallet is assumed to be "properly" loaded with the (CG) in the cargo's center directly above the center of the pallet. Each aircraft type has a specified number of pallet positions, each with a specified location, i.e. C-5s can carry 36 pallets at fixed locations in the aircraft while C-17s can carry 18 pallets.

SALP-TS requires two input files containing the problem's aircraft and the pallet information for the particular problem instance. For each aircraft, the *aircraft input file* contains: (1) a *unique* aircraft identification (ID) number, (2) aircraft type, and (3) aircraft ACL. The location and number of the pallet positions within a designated aircraft type is known and fixed. Pallets may be placed only in an available position.

The ACL is the maximum total weight loadable into the aircraft and depends upon the flight length, flight path conditions and available refueling aircraft (Defense Transportation Regulation 2004). The *planning* ACL is *only* a guideline for the weight which can be loaded into a particular aircraft. The aircraft types considered by SALP-TS with their planning and maximum ACLs are shown in Table 6.4.

For a particular SALP instance, there are three distinct possible combinations of available aircraft: (1) aircraft are of the same type with the same ACL; (2) aircraft are of the same type, but at least one has a different ACL; (3) aircraft are of different types. In the first two cases, the number of available pallet positions is constant for all aircraft. In the third case, this number varies among the available aircraft. The three combinations allow for different calculations of the *LB* on the number of required aircraft. This research assumes that the available aircraft will allow a feasible solution. To ensure this assumption, a solution to the problem is first generated using AALPS; the number of planes in the aircraft input file is equivalent to the number of planes required by AALPS which always produces a feasible solution.

In the AALPS *pallet input file*, the pallets are sorted by decreasing weight and, for each pallet, the file contains: (1) pallet ID number, (2) loaded pallet weight, (3) loaded pallet height.

A specific SALP-TS solution includes the fixed and known Total Number of Aircraft and Total Number of Pallets. The Number of Aircraft Used is a count of the aircraft used in that solution. The objective function (OF) value is the combination of OF values for each aircraft used in the solution as well as penalty fees for unused pallet positions. The Solution Array, associated with the aircraft pallet positions, and Big Bin List, pallets not assigned to an aircraft pallet position detail the actual location of the pallets. Initially, the Solution Array is empty and the Big Bin List contains all pallets.

The SALP-TS *solution representation* assigns pallets to specific positions in available aircraft. Each aircraft pallet position has a specific reference number associated with that aircraft. In single pallet row aircraft, numbering starts at the nose and proceeds sequentially to the tail, i.e. 1, 2, …, 6. In two pallet row aircraft, the pallets are labeled, by the US Air Force, as 1L, 1R, 2L, 2R and so on from the nose to the tail whereas SALP-TS labels them sequentially, i.e. 1, 2, 3, 4….

As a clarifying example, consider four C-130 aircraft available to transport 15 pallets. A possible solution for this problem is

(1, 4, 5, 6, 1, 2, 3)(2, 7, 8, 9, 0, 10, 11)(3, 12 13, 0, 0, 14, 15)(4, 0, 0, 0, 0, 0, 0)(0)

In each of the first four parenthetical subsets, the first number is the aircraft index and the remaining numbers are loaded pallet indices. A pallet index of 0 corresponds to an empty pallet position. In this example, aircraft 1 has pallet 4 in position 1, pallet 5 in the position 2, and so on. Aircraft 4 has no pallets loaded. The last set corresponds to the "Big Bin" (Harwig et al., 2006), i.e. "aircraft" 0 which has infinite capacity for weight and space. An empty Big Bin is represented by a single zero in the solution representation.

TABLE 6.4 Aircraft Type with Planning and Maximum ACLs

Aircraft Type #	Description	Planning ACL (lbs)	Maximum ACL (lbs)
0	C-130	25,000	40,000
1	C-17 (Logistics system, 18 pallets)	90,000	175,000
2	C-17 (Air drop system, 11 pallets)	90,000	175,000
3	C-5	150,000	291,000
4	KC-10 (17 pallets)	80,000	150,000
5	KC-10 (23 pallets)	80,000	150,000
6	C-141	46,000	70,000
7	KC-135E	30,000	40,000
8	KC-135R	30,000	40,000

An LB on the minimum number of aircraft required for a SALP is calculated by simultaneously considering both the total *planning* ACL and the total number of available pallet positions for the aircraft that will be used. The three distinct possible aircraft combinations, described above, require three separate methods for calculating the LB. When all aircraft are of the same type with the same ACL, the LB is calculated using the following two ratios rounded up to the next integer (as it is not possible to utilize a fraction of an aircraft):

1. ⌈(total number of pallets/number of pallet positions in the aircraft)⌉
2. ⌈(combined pallet weight/aircraft ACL)⌉

The LB is the maximum of the resulting two numbers.

When all aircraft are of the same type with differing ACLs, the LB is computed by sequentially removing aircraft by ascending ACL, starting with the smallest ACL, until a removal would cause either total pallet positions or total ACL to be insufficient.

When the set of available aircraft are of different types with different ACLs, SALP-TS sequentially removes, in ascending order, the still available aircraft with the smallest ratio of ACL to number of available pallet positions until any additional removals would cause either the total pallet positions or total ACL to become insufficient.

While it may not be possible to achieve the computed LB for particular SALPs, it provides a baseline for measuring the quality of solutions.

The SALP-TS OF to be minimized is composed of: (1) Big bin usage fees, (2) aircraft usage fees, (3) percent weight full penalties, (4) lateral CB penalties and (5) longitudinal CB penalties. The big bin usage fee, *BB*, for *each* pallet in the big bin forces SALP-TS to load all pallets onto available aircraft. The aircraft usage fee, *C*, for each aircraft used, drives SALP-TS to use fewer aircraft. The percent weight full penalty discourages aircraft ACL underloading and is computed as the product of the underloading penalty factor, λ_1, and multiplied times the sum of the squared underloadings for all aircraft.

Under certain circumstances, the planning ACL can be exceeded. ACL violations are penalized by the product of the overloading penalty factor, λ_2, and the sum of the squared overloadings. Since overloading is not preferred to under-loading, $\lambda_2 > \lambda_1$.

The CB of an aircraft is the location (measured in inches from an associated reference line) of the position of the center of balance of the cargo load. The CB is calculated by summing the moments of all pallets and dividing by the total weight of the cargo. The reference line is determined by aircraft type. The moment of each pallet is the product of the pallet weight and the distance from the reference line to the pallet's center of mass.

The lateral CB is only calculated for two pallet row aircraft; the reference is at the exact aircraft center line from nose to tail. For a C-5, the 108″ sides are parallel to the fuselage yielding pallet *CG* distances of 54″. In the C-17 and KC-10, the orthogonal pallet orientation yields *CG* distances of 42. The lateral CB contribution to the OF value is the sum of the squared CB locations for each aircraft multiplied by a penalty factor, λ_3.

In ordinary flight conditions, the lateral CB impact on the aircraft flight characteristic is very small; pilots can utilize the aircraft trim feature to enable the aircraft to fly straight and level without additional inputs. The flight contribution of the lateral CB can quickly become *significant* when turbulence or aircraft trim malfunctions occur. In these scenarios, the pilots are required to physically maintain control of the aircraft at all times. Any aid which can be rendered to the pilots in these times is invaluable. Properly loaded cargo (with a lateral CB near the aircraft center) allows pilots to more easily control the roll of the aircraft.

For the longitudinal CB, the *reference datum line* (RDL), a line at or near the nose of the aircraft, is specified for each aircraft type. Each pallet center has a *fuselage station* (FS) measurement (in inches) relative to the RDL which corresponds to the pallet's moment arm. The longitudinal CB is computed exactly like the lateral CB. Each aircraft has a target longitudinal CB location *at the current cargo weight* at which the aircraft obtains the best fuel consumption rate. The longitudinal CB has strict upper and lower bounds. A longitudinal CB outside these bounds will cause insurmountable aircraft flight control problems. Solutions are not feasible if the longitudinal CB location is outside these bounds. The computation of the longitudinal CB penalty is very similar to that of the lateral CB penalty. The squared deviation from the target value is computed for each aircraft and these values are summed and multiplied by a factor, λ_4. For the purposes of the search, if the longitudinal CB is outside of the allowable window, a much larger factor, λ_5 , is applied to ensure the solution is *much* less desirable.

The SALP-TS memory structure disallows return to the previous incumbent solution is forbidden for *tabu tenure* additional iterations. SALP-TS places an upper and lower bound on the tabu tenure. The tabu tenure is not allowed to drop below 4% and cannot exceed 20% of the total number of pallets. The tabu tenure is initialized at 12% of the total number of pallets. Including the adaptive tabu search (ATS) mechanism (Harwig et. al., 2006) enhances the effectiveness of SALP-TS by allowing the tabu tenure to dynamically change in response to the search. While respecting the stated bounds, an improving (disimproving) move causes a decrement (an increment) in the tabu tenure. This has the effect of intensifying the search in areas of improvement and diversifying when the solution has become worse.

The SALP-TS neighborhoods may be viewed as generating either fine or broad gauge moves. Fine gauge moves occur when a pallet is moved *within* a specific aircraft. In this case, a moved pallet is not allowed to return to the prior *position* for tabu tenure additional iterations. A broad gauge move occurs when one or more pallets are removed from aircraft. In this case, the pallet is prevented from returning to the prior *aircraft* for tabu tenure additional iterations. The tabu memory structure guides the search into areas with excellent solutions and helps escape areas of poor solutions. It is an essential part in the efficiency and effectiveness of SALP-TS.

SALP-TS tries to produce a quality initial solution (not necessarily feasible) at a small computational cost by attempting, for each aircraft used, to simultaneously maximize both the pallets loaded and the weight of loaded cargo. Weight is maximized when adding any additional available pallet would exceed an aircraft's ACL.

SALP-TS initially partitions the pallets into between one and ten groups which is user defined. The pallets are sorted by decreasing weight with the first group containing the heaviest pallets. SALP-TS attempts to equally divide the number of pallets between groups. If the total number of pallets precludes equal division, then the last (lightest) group will have any additional pallets.

In assigning pallets, SALP-TS selects the unfilled aircraft with the *largest* ratio of ACL to number of pallet positions, i.e. the aircraft that can carry the most weight per pallet position. Ties are broken by selecting the smallest aircraft index. Four aircraft states are possible during the initial solution process: (1) not maximized for either weight or space, (2) maximized for weight but not for space, (3) maximized for space but not for weight and (4) maximized for both space and weight.

For State 1 aircraft, SALP-TS begins with the heaviest nonempty group. That group's *head pallet*, the first available pallet which has not been already considered for the current aircraft, will be loaded if it will not violate the aircraft ACL. If the head pallet in any group cannot be loaded, SALP-TS advances to the next group.

If a State 1 iteration is completed with no additional pallets added, the aircraft must be in one of the other states. If the aircraft is in State 2, SALP-TS removes the aircraft's *heaviest* pallet, and places it into its original group. A pallet which is removed is not allowed to return for ten iterations. Next, SALP-TS selects the head pallet of the *lightest* nonempty group. If the pallet loading will not violate the aircraft ACL and the placement is not tabu, it is loaded. If the current head pallet is not loaded, it is marked as unavailable for the current aircraft. Regardless of whether the pallet is loaded, the group's new head pallet is considered. The algorithm does not stop at the end of a group, but rather it continues to the next group. This process is repeated until the aircraft is maximized according to weight, space, or both, or the end of the list of available pallets is reached.

When an aircraft is in State 3, SALP-TS removes the aircraft's *lightest* pallet and replaces it into its group and that pallet is not allowed to return for at least ten iterations. SALP-TS selects the heaviest available (nontabu) pallet that can be loaded without causing violation of the aircraft ACL. If no available pallet can be added, the pallet which was removed at the start of the iteration is reinserted and the aircraft is declared to be in State 4.

If an aircraft is in State 4, SALP-TS selects the next nonState 4 aircraft with the largest ratio of ACL to number of pallet positions and repeats the entire process. This continues until either all pallets have been loaded or no additional aircraft are available for loading.

SALP-TS could exhibit cyclic behavior in the initial solution process. To prevent cycling, one additional exit criteria is present. If an entire iteration is completed with no change in the number of pallets loaded onto the aircraft (i.e. no pallet is unloaded *or* a single pallet is unloaded and a single pallet is loaded), a variable labeled CycleCount is incremented. If there is a change in the number of loaded pallets, this variable is set to zero. If the CycleCount variable reaches a value of 10, SALP-TS declares the aircraft to be in state 4.

The SALP-TS initial solution technique ignores CB requirements and will usually produce sub-optimal and/or infeasible solutions. Any infeasibilities are corrected by SALP-TS's move neighborhoods.

Given an initial solution, SALP-TS performs four different types of pallet move neighborhoods. Each associated neighborhood serves a specific purpose in guiding the search to quality solutions. *Swap moves* cause two pallets to change positions and only involve pallets with different weights. *Insert moves* cause a pallet to change its current location by moving to an empty pallet position. The four neighborhoods are: (1) intra-aircraft insert-swap, (2) inter-aircraft insert-swap, (3) big bin to aircraft insert, and (4) unload an entire aircraft.

The intra-aircraft insert-swap neighborhood involves a *single* aircraft. Every possible swap of two pallets or insert of a pallet into an empty position is evaluated. The best nontabu move is performed. The lowest indexed pallet moved may not return to its previous position for tabu tenure future iterations.

The inter-aircraft insert-swap neighborhood involves *two* aircraft. Every possible swap and insert between two aircraft is evaluated and swaps of same weight pallets are disallowed. The best nontabu move is performed and returning a pallet to its donor aircraft is not allowed for tabu tenure future iterations.

The big bin to aircraft insert neighborhood removes *all* pallets from the big bin and inserts them *only* into an *empty* position in an aircraft. To execute this neighborhood, SALP-TS first selects the heaviest pallet in the big bin. SALP-TS then chooses an available aircraft with an empty pallet position possessing the greatest unsatisfied ACL and inserts the pallet into the aircraft's lowest indexed empty pallet position. The process repeats until the big bin is emptied. This neighborhood ignores both the ACL and CB constraints relying on SALP-TS to use other move neighborhoods to produce feasible solutions.

A primary goal of SALP-TS is to minimize the number of aircraft required to produce a feasible solution. A helpful diversification technique that assists in reaching this goal is to occasionally empty a selected aircraft. The donor aircraft is emptied in order of decreasing pallet weight and the pallets are inserted into empty pallet positions of available aircraft with the greatest weight capacity remaining. If every pallet position on other available aircraft, excluding the donor aircraft, is occupied, unloading ceases. Removing additional aircraft would result in insufficient pallet positions available to transport all pallets.

These four neighborhoods are strategically and dynamically employed to select the most appropriate neighborhood at each point in the search. The dynamic neighborhood selection process begins immediately after initial solution generation and is described in detail in Roesener (2006).

To demonstrate its efficiency and effectiveness, SALP-TS was directly compared with AALPS on the 12 different scenarios presented in Table 6.5.

The scenarios differ in the distribution of the pallet weights, the number of pallets to load, and the available aircraft types. "Equal distribution" pallet weights have an equal number of pallets weighing 2500, 4500, 7500, and 10,000 lbs. These were selected because AALPS is configured to allow the user to quickly select pallets with these weights. "Random weights" have pallets with weights uniformly distributed between 2500 and 10,000 lbs. Although SALP-TS can accommodate nine different configurations of six different military airlift aircraft, only the principle military strategic airlifters, C-17 and C-5 aircraft, are considered here.

TABLE 6.5 Aircraft Loading Scenarios

Pallet Weights	Number of Pallets	Aircraft Types	Scenario Number
Equal distribution	500	C-17	Scenario 1
Random weights	500	C-17	Scenario 2
Equal distribution	1000	C-17	Scenario 3
Random weights	1000	C-17	Scenario 4
Equal distribution	500	C-5	Scenario 5
Random weights	500	C-5	Scenario 6
Equal distribution	1000	C-5	Scenario 7
Random weights	1000	C-5	Scenario 8
Equal distribution	500	C-17 and C-5	Scenario 9
Random weights	500	C-17 and C-5	Scenario 10
Equal distribution	1000	C-17 and C-5	Scenario 11
Random weights	1000	C-17 and C-5	Scenario 12

The results of the algorithmic comparisons are presented in detail in Roesener (2006) and Roesener et al. (2008). Those results will be summarized here.

In terms of OF value, SALP-TS outperformed AALPS in every scenario demonstrating that the SALP-TS solutions not only required fewer aircraft, but also more closely achieved the target CB location on both longitudinal and lateral CB. In eight scenarios, SALP-TS produced feasible results which achieved the LB, while AALPS required more aircraft than the LB in every scenario. In 10 scenarios, SALP-TS produced feasible solutions requiring fewer aircraft than AALPS and required the same number of aircraft in two scenarios. Overall, SALP-TS produced feasible solutions which reduced the number of aircraft required by 0 to 8 aircraft with an average reduction of 2.33 aircraft. Trivially (Marginally) Infeasible solutions that allowed the ACL to be exceeded by no more than 1.5% (2.5%) on any aircraft were also studied but are not discussed here.

Another metric which demonstrates the effectiveness of SALP-TS solutions is the percentage of reduction in required aircraft of SALP-TS over AALPS. SALP-TS produced feasible solutions which reduced the required aircraft by 0–10% with an overall average of about 5%.

In fiscal year 2005, 61,796 hours of C-17 flight time were dedicated solely to channel (deployment) missions. If an average round trip flight requires 30 flight hours, then (61,796/30)≈2060 flights were planned. If the average cost of a C-17 channel mission is as previously described ($523,350), and if 5% or 103 of the 2060 C-17 flights could be eliminated without loss of throughput, a potential estimated savings of (103 * $523,350) = $53,905,050 per year could be obtained without loss of effectiveness. This estimate is only for *one aircraft type*; similar savings would be anticipated for the C-5. If one were to suppose the same number of C-5 flights were saved, an additional cost savings of (103 * $873,180) = $89,937,540 could be realized. The combined total savings from these two airframes would be approximately $53,905,050 + $89,937,540 = $143,842,590. This estimate also does not consider the maintenance costs or wear and tear on the aircraft. Another metric that is not considered in this estimate is the potential savings in fuel costs by loading aircraft with the intent of achieving the best fuel consumption rate, thereby requiring less fuel per flight.

6.7 Strategic Aircraft Routing

Once aircraft are loaded, they need to be routed. The current dynamic strategic environment places a premium on the effective (and to a lesser degree, the efficient) employment of air mobility assets. In testimony to the US Senate Armed Services Committee on operations in Afghanistan, General Thomas Franks, former commander of US Central Command, stated, "Strategic airlift remains key [to] current and future military operations. We are on a glidepath to expand our strategic airlift capabilities, and must remain committed to the task." (Franks 2002) The importance of this commitment is readily apparent when considering that in the 5 years since 9/11 (Hebert 2006; McNabb 2006):

- AMC flew more than 788,000 sorties, moved 7 million passengers and 3 million short tons of cargo, and offloaded 4.5 billion pounds of fuel. This is roughly equivalent to moving the entire city of Los Angeles.
- On 28 June 2006 AMC had 950 sorties scheduled worldwide; at that time the 6-month daily average was 825. This translates to more than one takeoff every 2 mins, 24 hours per day.
- Airlifted cargo recently surpassed the Berlin Airlift total from 1948 to 1949; current-operation flying hours (1.36 million) are as much as Berlin and Gulf War I hours combined.
- Since early 2003 the Aeromedical Evacuation system moved almost 75,000 patients to the US.

Strategic airlift is obviously an indispensable component of national military power. Like any other high-value and limited asset, it is imperative to carefully manage and effectively/efficiently utilize available strategic airlift resources so that US air mobility capabilities are equal to the daunting tasks of the present and future. Headquarters Air Mobility Command (HQ AMC) analysts and planners currently use automated tools ranging from simulations to spreadsheets in researching and analyzing strategic airlift problems (SAPs). For the most part, the tools employed are descriptive in nature and not designed to provide optimized flight schedules; aggregation is often required to ensure tractability and strategic-level outputs, which then forfeits operational-level detail.

First, note that SAPs are currently *not* clearly defined, mathematically-formulated problems, such as linear or dynamic programming problems. In general, SAPs are primarily concerned with *inter*theater airlift of personnel and cargo from the continental United States (CONUS) to a theater of operations, or from one theater to another. The United States Transportation Command's (USTRANSCOM's) tasking is to develop and direct the Joint Deployment and Distribution Enterprise to globally project national security capabilities, accurately sense the operating environment, provide end-to-end visibility, and rapidly respond to support joint logistics requirements. In addition to this, USTRANSCOM remains focused on providing global air, land, and sea transportation for the Department of Defense (DOD) in peace and war. The Air Force's AMC is the USTRANSCOM component providing global airlift, aerial refueling, and aeromedical evacuation support for both peacetime and wartime military deployment requirements (USTRANSCOM 2005).

At the outset of a contingency operation or war, USTRANSCOM receives personnel and equipment requirements from a Time Phased Force Deployment Document (TPFDD) developed by combatant forces and approved by the respective combatant commander. The TPFDD specifies such details as the personnel and equipment to be moved, their origin and destination locations, when they are available for loading at their origin, and when they are required to be delivered at their destination. Once USTRANSCOM determines the mode (air, land, or sea) of transport, requirements are then distributed to the appropriate USTRANSCOM subordinate commands, AMC for air transport, Surface Deployment and Distribution Command for surface transport, and Military Sealift Command for sea transport. AMC planners then select components from the Command's worldwide air mobility resources to construct an airlift network meeting the defined movement requirements. The network consists of aircraft, aircrews, airfields, support equipment, and personnel.

The number and type of aircraft employed and specific airfields available or used are a function not only of the TPFDD-derived airlift requirements, but of the geographic region of interest, in-place diplomatic agreements, and other relevant factors. Sources of the employed cargo and passenger aircraft may be either military or civilian; civilian airlift is obtained via the Civil Reserve Air Fleet (CRAF). The network's airfields include origin and destination airfields as specified in the TPFDD for each requirement. Origin airfields are termed *aerial ports of embarkation* (APOEs), and are where personnel and equipment are loaded onto aircraft. For most SAPs, APOEs will be located within CONUS where the bulk of US forces are stationed. Once transported, personnel and equipment are unloaded at destination airfields termed *aerial ports of debarkation* (APODs). APODs are generally located within or near the theater of operations. Since the large distances between APOEs and APODs generally preclude direct flights, AMC often incorporates a series of enroute bases (between APOE-APOD pairs) into the airlift network. These enroute bases may be used for refueling aircraft, changing and staging aircrews, and performing required aircraft maintenance.

Airlift crews are not assigned to a particular aircraft. Rather, aircraft move continuously through the network changing crews at enroute bases as necessary. A typical *mission* starts at an aircraft's permanent *home base* (HB) with a flight to an APOE, where some portion of the requirement (cargo and/or passengers) is loaded. The aircraft then passes through a series of enroute bases to the APOD where it is unloaded. The aircraft then flies to a *recovery base* (RB) for any required maintenance and crew change, and at this point, the aircraft returns to its home base. In less typical cases, the aircraft can skip the return to its home base and fly directly to another APOE to undertake a new mission.

Key SAP network flow constraints are the limited airfield resources for parking and servicing arriving aircraft. For example, these constraints include parking space, material handling equipment, available fuel, and support personnel. AMC planners aggregate such restrictions in a maximum-on-the-ground (MOG) value to capture the maximum number of aircraft that can be parked or serviced given certain operating conditions. The most common MOG value is the *working MOG*, which is a function of the parking MOG (number of aircraft that can simultaneously park at the airfield), aircraft maintenance capability, communications, fuel storage and transfer capacity. Because AMC currently owns a sufficient number of strategic airlift crews, their scheduling is not a limiting factor in solving some particular SAP. In other words, for planning/analysis purposes, crew scheduling may be successfully performed after all aircraft missions are scheduled. In real-time, real-world scenarios, crew positioning may be an issue requiring resolution.

The airlift network is the basic framework for the SAP. AMC uses the Air Mobility Operations Simulation (AMOS) to accomplish in-depth analysis of global mobility requirements and deployment capabilities. AMOS input files define the specific airlift network and problem instance for a given SAP. Major SAP components are requirements, aircraft, airlift operations, aerial refueling, cargo, locations (theaters, regions, airbases, waypoints), implicit/explicit resources, weather, routes, and other miscellaneous factors (Lambert 2004). Selected components are presented in more detail below.

6.7.1 Requirements

Key requirements are (1) type and quantity of cargo (labeled as outsize, oversize, and/or bulk), (2) commodity type (examples are *Navy_Marines* and *CSS_Other*), (3) number of personnel, (4) available to load date (ALD), (5) required Delivery date (RDD), (6) APOE and APOD locations, and (7) priority. The commodity type determines the fill efficiency achievable for a given aircraft.

Requirements are ordered according to ALD, RDD, and priority. The ALD and RDD define the requirement time window. Earliest arrival date (EAD) and latest arrival date (LAD) at the APOD must also be considered. Airbase MOG restrictions make "just in time" aircraft arrival preferable. Finally, a requirement is defined as being "on time" if it arrives on or prior to its RDD and its priority levels allow finer gauge requirement ordering.

Aircraft attributes include (1) aircraft type, (2) cruising speed, (3) cruising altitude, (4) fuel configuration (capacity, holding fuel, approach and landing fuel, taxi to takeoff fuel, and minimum landing fuel), (5) ACL by payload type, (6) trivial load definition, (7) standard times (ground times, onload time, offload time), (8) cargo compatibility (preferred, feasible, incompatible), and (9) body type (wide, narrow, small).

Cargo compatibility defines a specific aircraft's capability to move outsize/oversize/bulk cargo, and personnel, and is partitioned into preferred, feasible, and incompatible categories. (*Oversize cargo* is comprised of a single item that exceeds the usable dimensions of a 463L master pallet (104″×84″×96″) for military aircraft. A piece of *outsize cargo* is a single item that exceeds 1000″ in length, 117″ in width, and 105″ in height. *Bulk cargo* is cargo, including the 463L pallet that is within the usable dimensions of a 463L pallet and within the height and width requirements established by the cargo envelope of the particular model of aircraft.

An aircraft's body type determines its required parking and working MOG. To determine an aircraft's payload capability, this research uses a technique based on AMOS' "liquid ton modeling" method, which ignores the inherent spatial complexity of packing discrete cargo items and simply "pours" the cargo aboard the aircraft considering only the cargo's weight (Roesener, 2006; Roesener et al., 2008). (Note that AMOS also has a detailed loading algorithm which can be used in place of the "liquid ton modeling" method.) Key payload factors considered are the available cargo (commodity types), aircraft type, payload targets, and maximum cargo constraints for the given route.

Each airbase within a stipulated airlift network is assigned a unique identifier, often the International Civil Aviation Organization four-letter airport identifier code. Key airbase attributes include (1) latitude/longitude, (2) region, (3) MOG, (4) daily fuel capacity, (5) alternate landing base location, (6) compatible aircraft for landing (military or civilian; wide or narrow), and (7) permissions (e.g. operating hours).

Airbases may be home bases, APOEs, APODs, enroute or recovery bases, or some combination thereof. Each airbase is assigned to a geographic region and several bases may be contained within a region. A daily fuel capacity is apportioned to airlift. Alternate airfields are used if circumstances preclude landing at the primary airbase, but mission fuel requirements must provide enough fuel for flight to an alternate airbase. Recovery bases dictate where an aircraft recovers after offload.

For a given SAP, an AMOS input file defines the available APOE-APOD routes for that problem. A route defines possible specific route segments between source and sink regions that may be part of an aircraft's mission. Source regions are the mutually exclusive and exhaustive indexed geographic subregions (primarily in CONUS) from which requirements are transported. Sink regions are the mutually exclusive and exhaustive indexed geographic subregions (primarily outside of CONUS) to which requirements are transported. APOE and APOD bases are assigned to a specific source or sink region by virtue of their location within some selected region. These source and sink regions are joined together by the set of specified possible route segments, greatly reducing the total possible number of routes and complexity of the planning process. Figure 6.5 shows AMC-defined regions for some SAP. Navigation routes connect one numbered area with another. The navigation routes are either Category 1, which

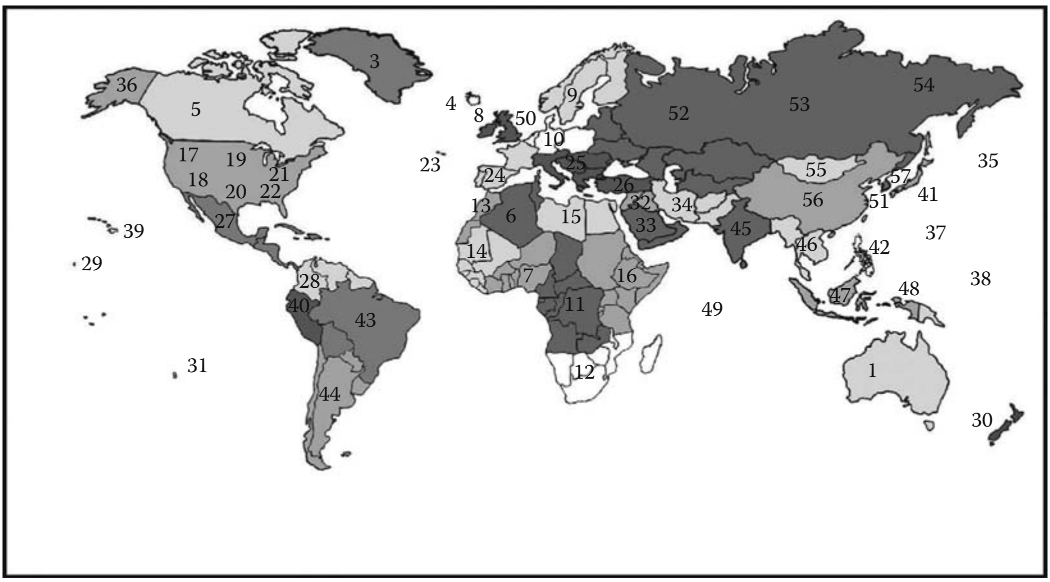


FIGURE 6.5 Region map.

traverse water and require additional fuel reserves, or Category 2. Category 1 routes require a minimum of 10% fuel reserve and one hour of cruise fuel be added to fuel requirements. Category 2 routes require no additional fuel reserves. Each plane's return trip is defined by a different planning group and route segment that may or may not be the reverse of the outbound trip. Finally, for each APOE-APOD route and aircraft type, a route's maximum payload weight is dependent upon the route's *critical leg* and the aircraft's characteristics.

In addition to the numerous inputs and constraints described above, solving a given SAP requires definition of a specific structured solution and objective function. In this research, a SAP solution requires the following decisions (consciously ignoring crews, as stated above):

- Detailed asset-level aircraft/requirement pairings
- Detailed aircraft missions flown
- Detailed time phasing of aircraft along routes to account for dynamically changing network characteristics (e.g. available MOG changes due to aircraft arrival and departure); this includes takeoff and landing times of each aircraft at APOEs, APODs, and at recovery and enroute bases

In solving a SAP, the "best" allocation of aircraft to requirements, and aircraft routing and scheduling is desired, so that late and nondeliveries are minimized and essential constraints are not violated.

A complete SAP *solution* is then given by the stipulated set of missions flown by the various employed aircraft, which then dictates the aircraft's payload, routing, and timings as it travels through the defined airlift network. A typical mission starts from the aircraft's home base and then moves through the airlift network servicing a particular APOE-APOD pair. The path through home base, APOE, possible en route base(s), APOD, recovery base and then back to home base is connected by route segments. Upon return to its home base, the aircraft undergoes maintenance and refueling and is then available for subsequent missions.

Mission timings are computed using AMOS input files defining times for the various stages of a mission: taxi to takeoff time, climb time, cruise time, approach and landing time, and ground time (onloading, offloading and recovery). The mission timing is computed by combining timings for all flight legs and ground times composing the mission.

A given SAP must consider multiple criteria, which may be viewed in a hierarchical manner; with the primary objectives of minimizing lateness of all requirements at the APODs, minimizing congestion at airfields in the airlift network, minimizing the number of missions flown, and minimizing the number of aircraft used.

Lambert (2004) and Lambert et al. (2007) define a SAP *objective function* based upon the following criteria: (1) undelivered cargo, (2) late cargo, (3) working MOG violations, (4) fuel violations (5) utilization target rate (UTE) violations, (6) trivial load violations, (7) number of trivial loads, (8) number of aircraft missions, (9) number of aerial refueling operations, and (10) objective function total score. The objective function's total score is computed using a linear weighted sum of the first nine criteria values where each criteria value is multiplied by a user-defined weight.

Undelivered and late cargo are measures of how well a solution meets RDDs for a given SAP instance. Undelivered cargo is defined as the total tons of all cargo not delivered to their respective APODs; late cargo is defined as the total tons of all cargo not delivered by their respective RDDs. Both undelivered cargo and late cargo are weighted according to the number of days the cargo is late. The weight multiplier for undelivered cargo is dependent upon a user-specified planning horizon and the weight multiplier for late cargo is the difference between the arrival time at the APOD and the cargo's RDD plus one.

Working MOG, fuel, and UTE rate violations are measures of a solution's feasibility. Working MOG violations are defined as the sum of the total number of MOG violations at each base at each time interval (the default time interval is 15 mins). Fuel violations are defined as the sum of the daily fuel shortages (in tons) at all bases under the current solution. A solution with a feasible UTE rate has no more aircraft in the air at any given time than that specified by the respective aircraft UTE rates. Thus, UTE rate violations are defined as the sum of the differences between the number of aircraft of each type in the air at any given time and the particular aircraft type's UTE rate. Working and fuel MOG violations indicate overly optimistic aircraft throughput in the airlift network under a given solution; UTE rate violations indicate overly optimistic aircraft operational rates. The default feasible settings for these measures are zero. Past AMC experience with AMOS-derived solutions *deemed feasible* for representative SAPs show these measures are rarely zero, especially for MOG and UTE measures. Thus, user-defined threshold values for these measures are permitted in this research to preclude overly constraining the airlift system.

The number of trivial loads, aircraft missions, and aerial refueling operations are simple counts, each with a default weight factor of 1. Trivial loads are loads whose weight falls below an aircraft's trivial load threshold. AMOS does not permit aircraft missions with trivial loads. Thus, depending on various user-defined rules, trivial loads are either aggregated at a central base until some weight threshold is reached and then delivered, or are simply not delivered at all. Trivial loads are permitted, but they are penalized. Missions with payloads less than the defined trivial load are counted and the trivial load violation is then defined as the total number of missions whose payload is less than the aircraft's trivial load threshold.

The number of missions is the current total number of assigned missions. The number of aerial refueling operations is the sum of all aerial refueling trips used in the current solution. AMC wishes to limit the number of aerial refueling operations supporting strategic airlift because its limited refueling assets are also dedicated to supporting fighter and bomber deployments. This is especially true in the surge phase of a major deployment.

Following Harder (2002), solution comparison is conducted by comparing each component of the objective function, one at a time, until either a "winner" is determined or all objective components have been compared and found equal. The order in which objective components are compared is as follows: the objective function total, undelivered cargo, late cargo, working MOG violation, UTE rate violation, fuel MOG violation, trivial load violation, number of aircraft missions, number of trivial loads, and the number of aerial refueling operations.

Lambert (2004) and Lambert et al. (2007) present an advanced dynamically adaptive TS approach to the SAP, SAP-TS. The SAP-TS solution representation is designed to compactly synthesize critical SAP aspects. A typical *mission* begins at an aircraft's permanent HB, followed by a possibly multileg route to an APOE where (some portion of) a requirement's cargo and/or passengers, are loaded. The aircraft then passes through a series of enroute bases to the APOD where it is unloaded. The aircraft then flies to a RB for any required maintenance and crew change. At this point, the aircraft usually returns to its home base before beginning another mission. An *arc* aggregates a set of airbases, i.e. *arcs* have an origin base, a route segment (possibly comprised of a series of intermediate bases) and a destination base. The origin and destination bases are the *end* bases of the arc.

Each mission moves through five distinct ordered points: HB, APOE, APOD, RB, and then back to HB. Thus, for a typical mission, an *arc* connects each pair of these bases. In SAP-TS, a SAP-TS *mission* is represented by a unique *mission index* and four *arcs* as follows: (Mission_Index, HB_APOE_Arc, APOE_APOD_Arc, APOD_RB_Arc, RB_HB_Arc). A SAP-TS mission contains all assigned payload, timings, and fuel consumption as well as any required routing information.

A SAP-TS *solution* is composed of the set of assigned missions. Since single aircraft can perform multiple missions, missions are ordered first by aircraft tail number and then in time order within tail number. Using an *arc* mission representation, Figure 6.6 shows two aircraft performing three missions. The number (in parentheses) next to each directed arc gives its *arc index*. Missions 1 and 2 are performed by aircraft X where both missions traverse the same arcs in the same order (arcs 4, 5, 6 and 7). Mission 3 is performed by aircraft Y and traverses arcs 8, 9, 10, and 11. All three missions begin at HB 1. Note that arcs 5 and 9 are likely to contain one or more enroute bases or crew staging bases. Furthermore, any arc (like *arc* 10) may contain some mandatory waypoint, i.e. a spatial location which must be flown through (traversed) prior to reaching a destination node.

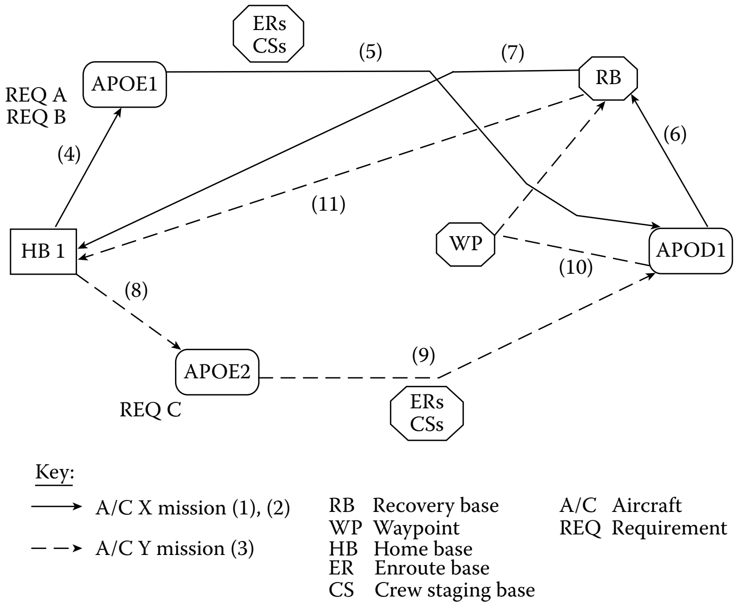


FIGURE 6.6 SAP representation using arcs.

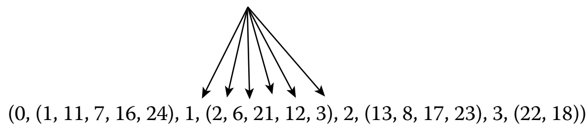


FIGURE 6.7 Possible NM1 inset points.

To represent each mission's content, the *arc indexed SAP-TS solution representation* lists the mission's ordered arc indices, *preceded* by the mission's index. Thus, the arc indexed SAP-TS solution representation comprising the three example missions, missions 1, 2 and 3, of Figure 6.7 is:

(1,4,5,6,7)(2,4,5,6,7)(3,8,9,10,11).

In the TS context it is *also* convenient to utilize a *APOE-APOD pair SAP-TS solution representation* to present the solution at a lesser level of detail, specifically, from the perspective of the APOE-APOD pairs. Note that while the APOE-APOD pair served by two or more aircraft missions can be the same, the HB APOE_Arc, APOE_APOD_Arc, APOD_RB_Arc, and the RB_HB_Arc all may be different. This less detailed APOE-APOD pair SAP-TS solution representation does *not* explicitly detail the arcs that make up each mission. In Figure 6.7, there are only two unique APOE-APOD pairs: APOE1-APOD1, which is assigned index 0, and APOE2-APOD1, which is assigned index 1. For each unique APOE-APOD pair, a time-ordered list of assigned missions is created. The APOE-APOD pair SAP-TS solution representation for missions 1, 2 and 3, depicted in Figure 6.7, would then be: (0, (1,2), 1, (3)).

As an additional example, consider the solution representation for the small problem detailed in Table 6.6. Table 6.6a presents the five aircraft (by unique tail number and by type) that are available to perform the missions required by this problem. Each aircraft is assigned five unique *possible* mission indices. (The number of possible mission indices should be dictated so that adequate missions are available to accomplish a solution to the particular SAP problem.) Table 6.6b presents the missions that are actually flown in the solution of the problem. The APOE-APOD pair SAP-TS solution representation for the solution given in Table 6.6b yields:

(0, (1, 11, 7, 16, 24), 1, (2, 6, 21, 12, 3), 2, (13, 8, 17, 23), 3, (22, 18)).

TABLE 6.6 A Solution to a Small SAP

(a) Aircraft and Missions Available

Tail Number	Aircraft Type	Available Mission Indices
1	C-5A	1–5
2	C-5A	6–10
3	WBC	11–15
4	WBP	16–20
5	C-17	21–25

(b) APOE-APOD Pairs and Assigned Missions

Pair Index	APOE-APOD Pair*	Assigned Missions
0	[KDOV, LTAC]	1, 11, 7, 16, 24
1	[KDOV, OKBK]	2, 6, 21, 12, 3
2	[KDMA, ETAR]	13, 8, 17, 23
3	[KBLV, RJBB]	22, 18

* KDOV = Dover Airforce Base, Dover, DE; LTAC = Ankara Esenboğa International Airport, Ankara, Turkey; OKBK = Kuwait International Airport, Al-Maqwa, near Kuwait City, Kuwait; KDMA = Davis-Monthan Air Force Base,Tucson, AZ; ETAR = Ramstein Air Base, Ramstein, Germany; KBLV = MidAmerica St. Louis Airport/Scott Air Force Base, Belleville, IL; RJBB = Kansai International Airport, Osaka, Japan.

Thus, aircraft 1 flies three missions (1,2,3), aircraft 2 flies three missions, aircraft 3 flies three missions, aircraft 4 flies three missions, and aircraft 5 flies four missions. Missions whose indices are absent are not flown, i.e. allocated *possible* missions 4, 5, 9, 10, 14, 15, 19, 20, and 25 are not required by the solution and are not flown. Using the two preceding SAP-TS solution representations, classical TS moves involving mission reordering and reassignment are easily applied in attempts to obtain better APOE-APOD pair mission assignments.

Lambert (2004) provides an *in-depth* description of the SAP-TS algorithm complete with psuedocode and examples. An overview of the SAP-TS algorithm will be given here. First, an initial solution is constructed incrementally using a simple one-pass method assigning aircraft missions to APOE-APOD pairs. Available missions are sorted according to the time available. The first mission for each aircraft is assigned an available time equal to the aircraft's original available time. Requirements are sorted and grouped by APOE-APOD pair, ALD, and aircraft compatibility. The initial solution construction method quickly yields a (typically infeasible) solution to the SAP but is suitable as a SAP-TS starting point. SAP-TS employs two alternating phases attempting to improve the current solution: (1) Strategic improvement (SI), and (2) LI. The SI phase broadly modifies the current solution by changing allocation of aircraft missions to APOE-APOD pairs. The *mission assignment/scheduler* heuristic then builds and schedules missions based upon those reallocations. The LI phase considers a set of missions with fixed APOE-APOD pairs, and then attempts to obtain improved loads, routings, and timings for those missions.

The *mission assignment/scheduler heuristic* operates in a manner similar to the initial solution construction method, with the major difference being that the mission assignment/scheduler heuristic obtains a solution using a given ordered assignment of aircraft missions to APOE-APOD pairs. Requirements are sorted by earliest ALD and then by RDD. The heuristic then attempts to sequentially load the next available requirement on the next available aircraft assigned to the APOE-APOD pair. If the current aircraft is not compatible with the current requirement or is not completely filled by the current requirement, other requirements available at the same APOE-APOD pair and on the same day are considered for loading. In this way, a solution is constructed based upon the ordered assignment of available missions at APOE-APOD pairs and prioritized cargo requirements.

Two additional measures drive the mission assignment/scheduler heuristic in the attempt to obtain better solutions. The first is associated with how *arcs* are selected during mission *construction* and the second with mission timings along arcs.

Arcs are selected using a rotating list for each set of route segments connecting an APOE-APOD pair. For example, suppose a C-5A's current mission requires a route from KDOV to OKBK, i.e. the APOE-APOD_Arc. The set of arcs connecting KDOV to OKBK under the current route plan are shown in Table 6.7. Since the C-5A is a military aircraft (type MIL), arcs 52 and 53 do not apply (as they require type CIV). Since this is an APOE-APOD route, it is desirable to maximize the possible payload. Maximum cargo values for arcs 54–66 range from a low of 77.31 tons to a high of 109.94 tons, with an average of 85.97 tons. Arc 66 has the greatest maximum cargo but also uses the most en route bases (three). Thus, the higher possible payload also strains the airlift network. If APOE-APOD arcs were selected solely on the basis of an arc's maximum cargo value, bottlenecks would likely develop along the arcs thus reducing throughput. For this reason, arcs are selected from the set of those available on a rotating basis. If the current mission was assigned arc 66, the next mission would be assigned arc 54.

The second improvement feature employed by the mission assignment/scheduler heuristic is a mechanism staggering aircraft flow through the airlift network. As arcs are selected, a running tally is kept of departure times from the arc's starting base. As an arc is employed in subsequent missions, a check determines if "too many" aircraft are starting down the arc at the "same time". "Too many" aircraft is defined as more aircraft moving down the arc than the minimum wide body aircraft working MOG (WBWMOG) for all the bases in that arc. WBWMOG is selected as the comparator because wide body aircraft slots at airfields are generally more limited than narrow body slots. Additionally, narrow body aircraft can often use narrow and wide body aircraft slots interchangeably. Aircraft are considered to be moving down the arc at the "same time" if they depart from the arc start base within a user-defined maximum ground time of each other (the default is 5 hours). Aircraft missions meeting this temporal criterion are counted and the total compared to the WBWMOG. If the WBWMOG is exceeded, the current mission's departure time from the arc start base is delayed by the ratio (maximum ground time/WBWMOG).

Critical decisions when solving SAPs include assigning aircraft to requirements, routing aircraft in assigned missions, and the detailed timings of aircraft movement through the airlift network. SAP-TS search *move neighborhoods* target problem domain aspects that encourage feasibility and improve

TABLE 6.7 Arcs Connecting KDOV to OKBK

Arc#	Seg#	AcTypes	C-5A Maximum Cargo (tons)	# En Route Bases
52	747	CIV	N/A	2
53	748	CIV	N/A	2
54	734	MIL	83.98	1
55	735	MIL	84.40	1
56	736	MIL	77.31	1
57	737	MIL	79.42	1
58	738	MIL	88.07	1
59	739	MIL	87.06	1
60	740	MIL	85.31	2
61	741	MIL	85.72	2
62	742	MIL	78.82	2
63	743	MIL	80.93	2
64	744	MIL	88.59	2
65	745	MIL	88.07	2
66	746	MIL	109.94	3

solution quality. Thus, given a set of APOE-APOD pairs and a set of available missions, favorable mission assignment perturbations improving the objective function are sought. Then, after changing current mission assignments, changes to individual missions are made seeking further improvement.

First we will consider SI phase neighborhoods:

1. The *new mission insert move neighborhood (NMI)* uses moves that insert currently unassigned missions into the solution, targeting APOE-APOD pairs with undelivered or late requirements. Undelivered requirements are considered first and sorted in descending order according to size. Late requirements are then considered and sorted in descending order according to the comparator ((tons of late cargo)×(days late)). Returning to the example in Table 6.6, suppose there is undelivered cargo for APOE-APOD pair 1 and there are late deliveries for APOE-APOD pair 2. In the interests of efficiency, only one unassigned mission per aircraft is considered (i.e., only missions 4, 9, 14, 19 and 25 are candidates for insertion as these are not already assigned in Table 6.6b and correspond to the available mission numbers for particular aircraft in Table 6.6a). New mission load feasibility is strictly enforced. Limiting possible inserts to those missions that address undelivered or late requirements further reduces NMI possibilities and focuses on inserts most likely to produce improved solutions.

   New missions addressing undelivered requirements should be inserted adjacent to current missions whose last requirement ALD is within three days of the undelivered requirement ALD or whose payload contains some portion of the undelivered requirement. Similarly, allowable insert locations addressing *late* requirements are the locations of those missions whose last ALD is within three days or whose payload contains some portion of the late requirement. These insert locations are chosen so that the new missions are assigned to the targeted requirement by the mission assignment/scheduler heuristic; the number of days is user defined.

   Given an unassigned mission to be inserted and an APOE-APOD pair with possibly previously assigned missions, the set of allowable insert locations for the unassigned mission is then determined. Again using Table 6.6's example, Figure 6.7 shows the six allowable insert points for unassigned mission 19 into the set of missions previously assigned to APOE-APOD pair 1. Of course, if another mission was selected for insert into the set of missions previously assigned to APOE-APOD pair 1, the allowable insert points could be different. In practical SAP instances, the number of allowable insert points is significantly less than the total number of possible insert points, thus greatly enhancing the efficiency of the search.
2. The *between pair swap neighborhood (BPS)* uses moves that swap aircraft missions previously involving *different* aircraft between two different APOE-APOD pairs. Note that swapping missions for the same aircraft is a null move. To identify poor mission assignments, the APOE-APOD pairs and the missions are sorted in descending order according to *mission quality value*. Mission quality value sums the late requirements value, (tons late)×(days late), and the unused aircraft capacity value, (aircraft ACL)−(current payload). APOE-APOD pair quality is then the sum of all assigned mission quality values and any undelivered cargo associated with the APOE-APOD pair. Unused aircraft capacity measures not only how well aircraft are loaded but also (indirectly) the current assigned missions' impact on MOG and UTE constraints; small loads result in more aircraft missions and hence a greater strain upon the airlift network.
3. *The within pair insert move neighborhood (WPI)* uses intensifying moves where assigned missions are moved to a different order position within their current APOE-APOD pair. APOE-APOD pairs are sorted in descending order by their contribution to the objective function, allowing the pairs contributing most to be investigated first. Changing the order of missions within an APOE-APOD pair allows the mission assignment/scheduler heuristic to discover better load efficiencies and improvements in temporally based objective values (working MOG, fuel violations, cargo lateness, and UTE rate). Based upon instance size, candidate mission insert points are restricted to be within an *insert move distance* from the current position. The insert move distance

is computed based on the average number of missions per APOE-APOD pair; WPI inserts are all possible inserts within the insert move distance to the left and right of the current position.

4. The *between pair insert move neighborhood* (BPI) uses moves to relocate a mission previously assigned to a particular APOE-APOD pair to another APOE-APOD pair. BPI moves are used in the search only to *escape* from a current unproductive region of the solution space by changing the cardinality of the current mission assignment structure. The associated reduction of missions in the donor APOE-APOD pair often results in a large undelivered penalty. Thus, BPI's use is curtailed to conditions where "escaping" is appropriate.

Next we consider LI phase neighborhoods:

1. The *within mission arc swap move (WMAS) neighborhood* moves target missions with MOG violations. Swapping *Arcs* in the current routing scheme can change the set of bases the aircraft traverses and thus reduce congestion at intermediate bases.
2. *The load reallocation (LR) move neighborhood* strives to increase fill efficiencies and reduce the total number of missions required. LR moves target undelivered cargo and those missions with payloads below the aircraft's trivial load threshold. If there is undelivered cargo, LR moves attempt to load that cargo on existing missions and assign cargo from aircraft with trivial loads to aircraft with excess cargo capacity.

   Spatial and temporal restrictions limit candidate missions for the LR move neighborhood. Candidate cargo-receiving missions must service the same APOE-APOD pair as that of the undelivered requirement or the trivial load mission. Additionally, a receiving mission's APOE arrival time must occur no than three days after the donor mission's ALD. The LR move distance is a temporal restriction akin to that used for the NMI move distance. Undelivered cargo moves are considered first and then moves for missions with trivial loads. If all requirements of a trivial load mission are removed, that mission is then deleted from the solution.
3. The *recover to APOE arc insert (RAI) move neighborhood* use specialized moves designed to reduce a requirement's lateness by altering an aircraft's routing on two sequential temporally adjacent missions. Specifically, RAI moves delete a RB to HB *arc* on the earlier mission and insert a RB to APOE *arc* based upon the APOE visited by the aircraft's next mission. In this way, the RB to HB time and HB to APOE time is replaced with the (shorter) RB to APOE time. RAI moves are constrained by the mission aircraft's return to base (RTB) time.
4. WMAS moves may not completely resolve MOG violations since violations occurring at arc endpoint bases are usually not affected by WMAS moves because only intermediate bases are exchanged. Arcs frequently have bases in common, so swapping an arc may not remove MOG violations at some intermediate bases. The *impose time delay (ITD) move neighborhood* is invoked to temporally stagger aircraft through the network in an effort to alleviate congestion and reduce the number of aircraft operating at any given time (i.e. reduce the UTE rate).

For MOG violations, bases are ordered by their respective total violation. Then for each base, the missions contributing to the violation are identified and sorted by *slack time*, which is defined as ((earliest mission requirement RDD)−(APOD mission arrival time)).

The intent is to first target missions that can be delayed without increasing lateness. Missions are considered in sorted order and an imposed mission delay is applied. This delay has the effect of staggering aircraft through the constraining base sufficiently to reduce MOG violations but without imposing too great a delay. For UTE rate violations, aircraft types are sorted for consideration using the respective aircraft type's UTE rate violation. Missions are sorted by aircraft type and then by slack within that type. Finally, missions are considered in sorted order and an imposed mission delay is applied.

A *memory structure* is constructed to efficiently and effectively control search, to prevent returns to recently visited solutions, and to allow search to escape from local optima. Moves to solutions whose attributes are restricted under the current memory structure are not allowed. There are two SAP-TS tabu memory structures which mirror the dichotomy in the SAP-TS move neighborhoods: (1) an SI phase memory structure, and (2) an LI phase memory structure. The SI phase memory structure assists with making broad changes to the current solution by changing the allocation of aircraft missions to APOE-APOD pairs; the LI phase memory structure assists in obtaining improved loads, routings, and timings for a set of missions with fixed APOE-APOD pairs.

The memory structure attributes for the SI phase consist of the APOE-APOD pair, mission number and original mission position in the APOE-APOD pair mission assignments. For insert moves (NMI, BPI, and WMI), only the original mission position of the single inserted mission is stored in the tabu memory structure. For swap moves (BPS), both original mission positions of the associated missions are stored. For example, suppose SAP-TS is at iteration 18, the current adaptive tabu tenure has a value of 7, and the current APOE-APOD pair to mission assignment is:

(0,(1,11,7,16,24),1,(2,6,21,12,3),2,(13,8,17,23),3,(22,18)).

If a BPS move is performed swapping mission 16 with 21, any move that either moves mission 16 back to APOE-APOD pair 0 at position 3 OR moves mission 21 back to APOE-APOD pair 1 at position 2 before iteration 25 is not allowed (the move is tabu). If a BPI move inserts mission 16 into APOE-APOD pair 2 at position 1, any move that returns mission 16 to position 3 in APOE-APOD pair 0 before iteration 25 is forbidden.

The memory structure for the LI phase mimics the mission representation: (MissionNode#, HBtoAPOEArc#, APOEtoAPODArc#, APODtoRB#, RBtoHBArc#), with a list of cardinality 5 (*msn*,*a1*,*a2*,*a3*,*a4*), where *msn* is the mission number and *a1* through *a4* are initialized to value 0. When any change to a mission occurs, the change is recorded in the memory structure element associated with the affected arc. For example, if an ITD move was used on mission 27 to delay the departure time at the aircraft's home base, a return to the original departure time for arc *a1* in mission 27's memory structure is not allowed for *tabu tenure* future iterations. Subsequent mission changes at later iterations for any mission at any arc are noted in the same way.

An LR move changes the payload for two missions, the *from mission* and *to mission*. Neither payload may be returned to the original payload for *tabu tenure* future iterations.

WMAS and RAI moves represent changes in the arcs used by the mission. The original changed arc cannot be reused by the changed mission for *tabu tenure* future iterations. In this way, move attributes affecting specific portions of the mission, payload, or routing are recorded to prevent returning to previously visited solutions during the extent of the *tabu tenure*. That said, SAP-TS can implement a tabu move if it satisfies the default aspiration criterion, i.e. the resulting objective function value is superior to any previous solution's objective function value.

The *SAP-TS Move Manager* is the heart of SAP-TS and determines the appropriate phase, search context, i.e. diversify, intensify or escape, and move neighborhood to apply to the current solution. Move neighborhoods are dynamically selected based upon search history. SAP-TS uses several mechanisms to determine when either the SI or the LI neighborhood phases should be employed. SI moves are more computationally expensive because they invoke the mission assignment/scheduler heuristic. LI moves are much simpler as they update only directly affected missions. For this reason, the SI maximum neighborhood size is smaller than the LI maximum neighborhood size.

Figure 6.8 gives an overview of how the search phase and context are determined. Starting at the figure's top, presume a new incumbent solution has just been obtained. *Escape* is employed if the search history indicates that search has stagnated and more radical steps are necessary to ensure escape from the current search space. Escape conditions are met, if, within the previous 20 iterations, the number of new objective function values within one percent of the previous objective function values exceeds 10. If escape is invoked, the phase is set to SI, the neighborhood to BPI, and three escape iterations are executed.

If the last move was an improving move, the search context is set to "intensify" and the neighborhood from the previous iteration is used in the current iteration in an attempt to more densely search in the vicinity of the current solution. If the last move was not an improving move, the search algorithm

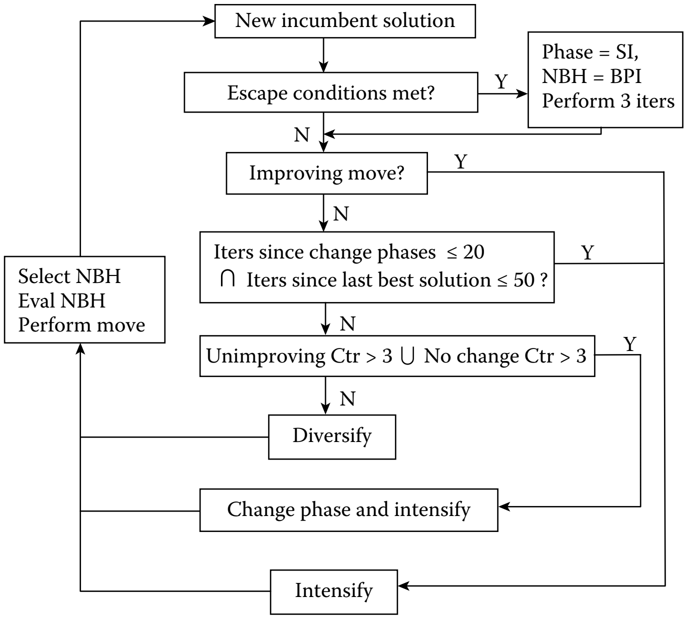


FIGURE 6.8 Determination of neighborhood phase and search context.

determines whether *both* (a), less than or equal to 20 iterations have passed since the last phase change, *and* (b), less than or equal to 50 iterations have passed since a new best ever solution has been found. If both conditions (a) and (b) are satisfied, the search context is set to "intensify" and the neighborhood from the previous iteration is used. However, if one or both of conditions (a) and (b) are false, the search algorithm determines if *either* condition (c), the unimproving move counter exceeds 3, *or* condition (d), the no change counter exceeds 3. If either condition is true, the search context is set to "intensify" and the search phase is "toggled", i.e. if the search phase was SI, it is changed to LI and vice versa.

If neither condition (c) nor (d) is true, the search context is set to "diversify" and the current iteration's neighborhood will differ from the previous iteration's, thus causing the search to move away from the current solution. The specific way that the neighborhoods will be changed from iteration to iteration depends on the search phase and is detailed later.

When the initial incumbent solution is constructed at the start of the first iteration, the search phase is initialized to SI and the algorithm, as described above, sets the search context to intensify. When the search context is "intensify", one of two neighborhoods are selected. For SI, if the current solution has undelivered or late cargo, the selected neighborhood is either NMI or BPS; otherwise, the selected neighborhood is either BPS or WMI. For LI, if the current solution has undelivered or late cargo, the selected neighborhood is either LR or BPS; otherwise, the selected neighborhood is either BPS or WMI.

When the search context changes to diversify, the new neighborhood is determined based on the search phase, the previous neighborhood, and the current objective function status. For SI, if the current solution has undelivered or late cargo, the selected neighborhood alternates between NMI and BPS. For LI, the neighborhoods are employed in the following order: LR, WMAS, ITD, and RAI. Neighborhood evaluation is terminated when either all neighbors are considered or when a maximum of 1000 move evaluations are performed. This methodology embodies the context of a *functional candidate list* which limits the size of the neighborhood to be evaluated while ensuring the most promising neighbors are considered first.

The various settings for the numeric parameters mentioned above are based on the authors' knowledge of the problem, the algorithm domain and structure and upon empirical experimentation. These settings were identical for all runs used in generating the results.

After the initial solution is constructed, the search phase is set to SI and search context is set to intensify. If undelivered cargo exists, the neighborhood is set to NMI. Otherwise, the neighborhood is set to BPS. Once the neighborhood is set, a stipulated number of moves are evaluated and the best nontabu or aspiring move is selected. The selected move is performed and the solution updated. If the move is disimproving, both the adaptive tabu tenure and the disimproving move counter are incremented. If the move is improving, the disimproving and "no change in objective function value" counters are reset to zero. If a new best solution is found, the tabu tenure is reset to the default minimum. The move manager is then invoked to determine the appropriate phase, context, and neighborhood to employ based upon the new solution and search history. This process repeats until either the maximum number of iterations has completed or the maximum search time has elapsed.

After the standard search phase (as described earlier) is completed, SAP-TS performs a concluding elite search phase beginning with the best solution yet found. Elite iteration neighborhoods are restricted to the mission improvement phase neighborhoods (LR, WMAS, ITD, RAI) to intensify search around the elite solution.

SAP-TS was applied to five different SAP scenarios. AMC provided these scenarios to represent operationally realistic SAP problem instances, and to provide a sound basis with which to investigate the efficacy of SAP-TS across a variety of scenarios typically encountered by AMC.

AMC designed each scenario to emphasize different aspects typical of SAP instances. In particular, feasibility constraints and problem sizes were varied in order to provide a robust assessment of SAP-TS. To discern between different scenarios, the following key problem parameters are identified: number of aircraft, number of unique APOE-APOD pairs, total cargo (tons), total number of personnel, and scenario emphasis. These scenarios and their key features are summarized in Table 6.8.

Several measures were used in generating *comparative results* for SAP-TS and AMOS across the five scenarios. First, the SAP-TS objective function was used to provide a *holistic* measure of results. Second, the *makespan* or closure time for the respective solutions was computed to measure the time required for all requirements to reach their respective APODs under the two models. Finally, the daily closed cumulative tonnage was tabulated to determine how required delivery times were being met.

SAP-TS executed each scenario for a total of 15 hours on a Dell Workstation equipped with a 2.4 GHz XEON CPU and 2 GB of RAM. The standard search phase was limited to 12 hours. The number of elite search iterations performed is either 800 or the number of iterations performed in 3 hours, whichever is less. In executing any advanced TS method like SAP-TS, it is very easy to compile a set of excellent solutions of a specified cardinality at no meaningful additional cost in computational effort. In SAP-TS, this was done both for preparation for the elite search phase and during the elite search phase. This provides a possible set of solutions from which a decision maker might select a solution deemed better based on qualitative motivations not included explicitly in the SAP-TS search. In each of the scenarios discussed below, an ensemble of competing solutions to the best one reported here were found. In the interest of required brevity, only the best solutions are presented here.

As stated earlier, AMOS executes on Windows-based systems. AMC analysts use a number of such systems on a day-to-day basis, many of which vary in processor speed and memory size. However, it is

TABLE 6.8 Scenario Characteristics

	Problem Parameters				
Scenario	Number of Aircraft	Number of Unique APOE-APOD Pairs	Total Cargo (tons)	Total PAX	Scenario Emphasis
Scenario 1	297	56	7647	14,769	(UTE/LATE)
Scenario 2	124	45	15,482	10,576	(LATE/MOG/FUEL)
Scenario 3	297	95	13,777	16,683	(MOG)
Scenario 4	165	12	7647	14,769	(LATE)
Scenario 5	70	12	3786	9769	(Typical)

TABLE 6.9 AMOS and SAP-TS Results for Scenarios 1–5

	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
Criterion	AMOS	SAP-TS	AMOS	SAP-TS	AMOS	SAP-TS	AMOS	SAP-TS	AMOS	SAP-TS
Undelivered cargo	144.40	0.00	0.80	0.00	0.00	0.00	4.95	0.00	0.00	0.00
Late cargo	7005.83	69.61	9992.05	8978.00	1595.08	135.20	20,974.74	749.55	762.97	16.24
Working MOG	163.00	114.00	1038.00	1033.00	51.00	41.00	578.00	314.00	58.00	16.00
Fuel MOG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# Missions	240.00	248.00	425.00	389.00	116.00	228.00	577.00	403.00	116.00	138.00
# Trivial loads	56.00	12.00	88.00	34.00	19.00	2.00	239.00	22.00	19.00	3.00
Trivial load penalty	507.89	27.00	805.28	348.32	136.92	1.51	1837.79	71.70	136.92	9.50
# Aerial refueling missions	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UTE rate violation	4237.00	141.00	30,096.0	5069.00	1623.00	11.00	12531.0	2718.00	0.00	0.00
Total objective value	12,354.1	611.61	42,445.1	15,851.3	3541.00	418.71	36,742.5	4278.2	1092.9	182.74
Makespan	31.52	26.99	18.40	18.82	29.02	28.62	40.47	32.35	30.85	26.30

fair to say that a typical AMOS scenario executes in 15–30 mins. This would imply that a knowledgeable AMOS user would have, at most, only somewhere between 30 and 60 labor intensive simulation runs and analyses to attempt to reach results comparable to SAP-TS. Perhaps a hybrid methodology that joins SAP-TS and AMOS would be a fruitful arena for further investigation.

AMOS and SAP-TS results for Scenarios 1 through 5 are summarized in Table 6.9. SAP-TS performance dominates AMOS' across the five study scenarios, as measured by the total objective function and key components such as late/undelivered cargo and feasibility. The best SAP-TS solutions achieve reductions in AMOS' objective value ranging from 63% to 95%. In all scenarios save one (Scenario 2), SAP-TS closes all requirements before AMOS. For Scenarios 1, 4 and 5, the best SAP-TS solutions close almost one week before AMOS; in Scenarios 2 and 3, the difference in closure times is less than half a day. Further details and analysis of experimental results are described in Lambert (2004).

Although SAP-TS performance appears impressive, several important caveats must be noted. Although AMOS maintains sufficient operational-level detail to potentially enable detailed and efficient routing/scheduling of strategic airlift resources, AMC typically does not use that information in scenario analysis. AMOS is a *simulation* used to examine potential futures arising from changes in key input parameters (e.g. fleet size, aerial refueling availability, etc.) AMC analysts typically use AMOS to support decision makers through long-term studies; many taskings result in a very short (e.g. 24 hours) turn-around. None of these uses require producing efficient solutions but do require quick execution.

Thus, readers should not view these results in light of typical computational experiments directly comparing algorithmic performance over well-chosen problem domain instances, with clear and unbiased metrics, *as AMOS and SAP-TS are designed for different purposes and with different desired end-states*. AMOS is a well-understood simulation capable of solving large SAPs and was very useful as a point of comparison in extending TS methodologies to solve these challenging COP.

6.8 Aerial Refueling

Another important aspect of the strategic deployment problem is air refueling, the in-flight transfer of fuel between tanker and receiver aircraft which provides rapid response, increased range, and extended airborne operations for all aircraft. The capability to perform air refueling makes the United States the dominant air power in the world today. This capability, coupled with the ability to efficiently employ air refueling assets, is essential for the US to maintain this dominance (AFDD 2-6.2 1999). The AMC of the USAF is responsible for determining the use of the US tanker fleet to meet the air refueling needs of the Air Force, Navy, Marines, US allies, and coalition partners. Part of AMC's planning encompasses the intertheater movement of forces from the US to areas around the globe. This "deployment" of forces and its accompanied air refueling requirement is known as the AFRP. Given a deployment scenario, examples of overview questions that require answers are:

- How many tankers are required to meet the air refueling requirements?
- How quickly can all the receivers be deployed to their final destinations?
- How far do the tankers and receiver aircraft have to travel?
- How much fuel is burned by both tankers and receiver aircraft?

In order to meaningfully answer upper level questions like these, as well as more probing questions relating to the efficiency of planning operations, a great many detailed operational aspects must be addressed.

We assume that the following information is given:

- A known set of tankers and their associated original beddown (starting) bases.
- A known set of receiver aircraft, each with an initial departure base and a final destination base, where one or more receiver aircraft are aggregated to form receiver groups (RG's). Each RG has a stipulated desired earliest arrival time and latest arrival time .

- A known set of RG starting and ending bases and tanker beddown bases.
- A known set of flight characteristics for each aircraft including flight speed, altitude, take-off weight, fuel capacity, and fuel-burn rates.
- A known set of tanker specific characteristics including fuel-offload capacity and fuel-offload rates.

Waypoints (WPTs) define the physical locations and start times where each refueling of an RG will take place. For a given deployment, the following decisions compose the solution to the AFRP.

- The waypoints (WPTs).
- The tanker(s) that will serve each WPT.
- How much fuel the assigned tanker(s) should deliver to a WPT.

We assume that the decision maker has the authority to (a) stipulate the departure times of all RGs and tankers and (b) to require both tankers and RGs to "orbit" at specified locations to satisfy WPT requirements in terms of timing and location. These two assumptions allow an AFRP solution to satisfy a number of constraints while exhibiting a number of desirable qualities to include deconflicting schedules and reducing the number of tankers required. The research documented in Wiley (2001) and Barnes et al. (2004) concentrates on determining the "best" assignment and ordering of WPTs with tankers while generating a flyable schedule for the tankers and RGs alike.

The AFRP objective function, as implemented by Wiley (2001) and Barnes et al. (2004) is multicriteria and hierarchical. The hierarchical ordering of the associated criteria is subject to redefinition in accordance with perceived mission priorities. The following criteria, in the order given, defined the specific hierarchical objective function addressed.

Minimize:

1. The number of unescorted RGs requiring escort between WPTs
2. The number of WPTs not serviced by a tanker
3. The number of WPTs serviced out of order for a RG's flight path
4. The amount of required fuel not available
5. The amount of time spent by RGs and tankers in 'orbit' at WPTs
6. The amount of RG late arrival time, i.e. where one or more RGs arrive later than a desired latest time
7. The number of tankers used
8. The amount of tanker mission time required
9. The total distance flown by tankers
10. The amount of fuel used by tankers
11. The amount of fuel off-loaded by tankers
12. The amount of fuel used by RGs

For any pair of solutions, superiority is determined by strict hierarchical criteria comparison starting at index 1 and traversing through increasing indices. At any index, a lower criterion value (bounded below by zero) implies a better solution and the comparison process is stopped. Two solutions are equivalent only if all 12 criteria are equivalent.

The AFRP solution is constrained by a large number of limiting factors. The safety of the crews and aircraft associated with the deployment is dominant, i.e. no tanker or receiver aircraft should have to divert from its flight path due to lack of fuel. Many other constraints must also be satisfied. For example, a tanker has limited fuel capacity and its crew has flight duration restrictions which affect the crew-tanker ability to travel long distances and to provide fuel. Certain bases have limited capacity for resident aircraft (known as *maximum on ground* or MOG). In the scenarios considered in this research, the important constraint of MOG was not a consideration, but could be addressed in future analysis. A tanker requires time to fly between any two locations and time to perform midair refueling. Hence, all tanker WPT assignments must be limited to those where the tanker is physically capable of being present at the specified WPT time.

The AFRP is unique, complicated, and extremely difficult to model and solve when viewed in its full practical context. Perhaps the most closely associated classical problem is a variation of the multi-vehicle, multi-depot vehicle routing problem (VRP).

In the AFRP, we have finite capacity heterogeneous vehicles located at multiple depots (with route length/route duration upper bounds) that are required to deliver product to customers (such deliveries require a finite service time). In the notation of Carlton (1995), the AFRP is a variation of problem type $\alpha = (MV\overline{H}, MD, VRP, RL)$, where $MV\overline{H}.\equiv$ multi-vehicle heterogeneous, $MD.\equiv$ multi-depot, $VRP \equiv$ vehicle routing problem, and $RL.\equiv$ route length, which is known to be NP-hard (Gendreau et al. 1997).

However, there are several additional considerations present in the AFRP that are not present in problem α. These are:

1. In problem α, the customers' locations are fixed in space, requiring only that the customers be partitioned among the vehicles and that each vehicle's customers be ordered for service. Further, the amount of product to be delivered to each customer is a known amount and there is a single delivery to any customer. Finally, the route length restriction is given only in terms of a total travel distance that may not be exceeded. Problem α has no explicit accounting for the timing of events. As in problem α, we must stipulate the responsible vehicle (tanker) and the ordering of any delivery. In addition, for all RGs, we must also stipulate the spatial location (longitude, latitude, and altitude in 3D space) and start time of each fuel delivery and the number of possibly multiple deliveries and the amount of product (fuel) to be provided in each delivery.
2. All customers must be supplied with fuel in a timely manner that will assure that no receiving aircraft has its available fuel fall below a prespecified "minimal reserve".
3. Directly associated with the WPT decisions are the decisions on the takeoff time of each RG and the possibly multiple takeoff times of each tanker.

The primary objective was to develop methods for producing an "ensemble" of excellent solutions to any instance of the AFRP. To achieve this objective, a *group theoretic TS* (GTTS) approach was used. GTTS makes use of adaptive TS (Dell'Amico and Trubian 1993, Glover and Laguna 1997) to dynamically update the memory structures, promote diversification, and remove the need to empirically "tune" a static tabu tenure. GTTS represents a solution to the AFRP as an element from the symmetric group on n letters, S_n (Isaacs 1994), and creates move neighborhoods that can be represented by S_n acting under conjugation or multiplication upon itself. To address the issue of reusability and portability, the GTTS approach was implemented using the Java™ programming language. Our implementation makes extensive use of Wiley's (2001) Java™ class library for S_n. Other goals were to investigate the effects of selecting different move neighborhoods both from a static and dynamic context. In order to effectively construct a GTTS for the AFRP, five primary types of *implementation* objects are defined and used: (1) *locations*; (2) *aircraft*, (3) *RGs*; (4) *tankers*; and (5) *nodes.*

These objects store detailed information beyond the partitioning and ordering captured by S_n. The connection between S_n and these objects occurs through a mapping of node labels to the integer set. Based on the placement of these node labels (a permutation of S_n), the information, described in the following paragraphs, associated with tankers, RGs, and nodes will be determined.

A solution of the AFRP requires detailed information about (1) the physical locations of bases and WPTs, (2) the flight characteristics of aircraft, (3) the assignment of aircraft to RGs and the stipulation of the flight path that a RG will follow, and (4) the number of tankers available for refueling activities and their associated beddown bases. In addition, information about the actual AFRP solution is stored in the tanker and RG objects. This is accomplished in the following ways: (information not germane to a particular object is left blank).

Locations (bases, WPTs): (a) unique ID, (b) code name, (c) coordinates, (d) MOG, (e) whether over open water (after all location objects are created, a symmetric distance matrix is generated using the great circle' distances associated with every pair of locations (Capehart 2000)).
Aircraft (light receivers, heavy receivers, tankers): (a) unique ID, (b) airframe type (light, heavy, tanker), (c) nominal true air speed, (d) total fuel capacity (e) fuel burn characteristics (required fuel reserve, nominal take-off fuel, nominal fuel flow coefficient, nominal altitude, empty weight, and nominal load weight).
Receiver groups: (a) unique ID, (b) total fuel upload requirement, (c)list of aircraft IDs in the RG (determines escort requirement), (d) flight path information (start and end base IDs, RG flight characteristics, earliest departure time (EDT) and latest departure time (LDT)), (e) RG solution attributes

- (e1) List of flight path WPT nodes
- (e2) Arrival, service, and departure times at each WPT node
- (e3) Amount of fuel burned between WPT nodes
- (e4) Amount of fuel required to completely refuel each member of the RG
- (e5) Start and finish times

Tankers: (a) unique ID (b) off-load capacity, (c) tanker solution attributes (blank unless tanker used)

- (c1) WPT nodes served and associated times (arrival, service, and departure)
- (c2) fuel burned and fuel offloaded for each WPT served
- (c3) tanker start and finish times for deployment

Nodes (tanker, WPT, RTB): (a) unique ID (implies node type), (b) spatial location and fuel requirement, (c) RG assigned, (d) precedence relations with other nodes, (e) whether linked to another node for escort duty, (f) tanker assigned, (g) the fuel demand.

Since the node objects are used to represent disparate entities, some additional remarks are appropriate. A tanker node uses only field (a). Creating a feasible solution (where no tanker runs out of fuel), may require some of the tankers to return one or more times to a base capable of tanker maintenance and refueling. At such a base, completion of these operations allows the tanker to continue servicing WPTs. To allow for this, tanker return to base (RTB) nodes are created for each active tanker base (as needed). RTB nodes are distinguished by zero fuel demand (from a tanker) and the lack of an assigned RG. Created only if needed, an RTB node will possess a unique ID (field (a)), its spatial location (field (b)) will correspond to the associated refueling base, and the tanker ID of the associated tanker will occupy field (f). (All other fields are blank for an RTB node.)

A WPT node uses all seven fields. Field (d) will contain the information to assure that WPT nodes are visited in precedence (temporal) order along the RG's flight path. If a consecutive pair of WPT nodes within a RG's flight path are both over water and the RG has escort requirements, field (e) will indicate that the flight "leg" between the WPT pair requires escort by a tanker. Field (f) remains blank until a tanker is assigned to the WPT node as part of a generated solution to the AFRP. Field (g) will contain the RG's fuel demand at that WPT.

In solving a particular AFRP, the tanker nodes are generated first and are sequentially assigned indices starting at index 0. The WPT nodes are then created in the ascending order by RG ID and then by ascending order of associated WPT ID's within the RG flight path. For each WPT, its WPT nodes are indexed sequentially starting with the next available index. For every WPT along each RG's flight path, either 1, 2, or 3 WPT nodes are generated.

Only one WPT node is required if both of the following conditions hold true: (a) the RG's fuel demand at that WPT can be satisfied by a single tanker and (b) the RG does not require an escort to its next WPT.

Two WPT nodes are required if either of the following two sets of conditions are true: (a) the RG's fuel demand at that WPT requires multiple tankers and the RG does not require an escort to its next WPT and (b) the RG's fuel demand at that WPT can be satisfied by a single tanker and the RG requires an escort to its next WPT.

If condition (a) holds, the WPT is represented by two WPT nodes where each is assigned one-half of the original WPT's fuel demand. With this representation, a different tanker can be assigned to each WPT node, jointly satisfying the total WPT demand.

If condition (b) holds, the WPT is represented by two WPT nodes. The first node is assigned all of the WPT demand and the other is assigned a demand of zero. These WPT node creations and assignments are made so that one tanker can perform the refueling function at the WPT and, if appropriate, another tanker can perform the escort duty to the next WPT. (The escorting tanker will provide the refueling function at the next WPT in the RG's flight path.)

Three WPT nodes are required if the RG's fuel demand at that WPT requires multiple tankers and the RG requires escort to the next WPT. In this case, two of the three WPT nodes serve to allow the required refueling by two different tankers, and it is possible that another tanker will assume the escort duty to the next WPT by being assigned to the third WPT node. When the WPT nodes are created, the fuel burned between a RG's adjacent WPTs is calculated and becomes the fuel demand for the latter WPT. RTB nodes are generated as needed and are assigned the next available index. Once the AFRP node objects have been created, they are used in the GTTS implementation using the OpenTS framework of Harder (2000).

A solution in the *AFRP-GTTS* approach is represented by a permutation, $s \in S_n$, where n is the cardinality of the set of node objects defined for the current instance of the AFRP being solved. Each letter in a solution, s, corresponds to the index of an AFRP node object and each solution is stored as a symmetric group object. Each solution, s, embodies methods for creating moves, for performing conjugation and multiplication of elements of S_n and the current tanker assignments are implied in the solution's cyclic structure and the node indexing scheme described above. For example, the cycle, (0,15,16,17), would be interpreted as "Tanker 0 flies to (services) AFRP nodes 15, 16, and 17 and then returns home". All other tanker assignments are also represented with additional cycles. Indeed, all cycles in a solution will contain a single tanker node, which, by convention, will be placed in the first position of each cycle in the permutation representation.

Since S_n acts upon itself, all AFRP moves are also represented as permutations in S_n and move permutations act directly on solution permutations to produce different AFRP solutions. At each iteration, a defined neighborhood of eligible moves is evaluated. The neighborhoods generated by the AFRP are move-based neighborhoods rather than the more traditional solution-based neighborhoods. A solution-based neighborhood considers solutions reachable by one move. Rather than storing solutions, the moves that transform the current solution to a neighboring solution may be stored. The move neighborhoods for the AFRP are composed of permutations in S_n and the move-based neighborhood is a natural extension of this concept.

After a candidate list (Glover and Laguna 1997) of moves is stipulated, the solution resulting from each move is evaluated relative to the hierarchical objective function. This evaluation automatically performs any operations required to deconflict the schedules of the tankers and RGs in the new solution. Any solution that is the best solution found so far (within the current iteration) causes a check on the tabu status of the associated move. If the move is not tabu, that move is stored as the iteration's current best move found. Once the best available move is selected, that move is implemented and recorded within the tabu structure and the TS procedure begins a new iteration.

When an aspiration criterion is used, a move satisfying the associated criterion will be accepted regardless of its tabu status. An aspiration criterion object allows the definition of a criterion for any particular solution/move combination. The most common aspiration criterion used states that when a new solution is found that is superior to any found earlier in the search, the new solution is accepted regardless of its tabu status.

For the current discussion, we assume that the WPT locations for all RGs have been supplied by an external source. This case prevails under previous AFRP solution techniques. (We present an adaptive TS approach for the WPT location problem later in the discussion.)

The initial solution created for the AFRP assigns all WPT nodes to the first tanker (tanker 0). Usually this will produce a highly infeasible solution (i.e. the tanker's capacity will be exceeded). If a problem had 15 tankers and 32 WPT nodes, the initial solution would be (0,15,16,17,…,44,45,46). To overcome this infeasibility, the *tanker insert* (TKI) move neighborhood is used to insert unused tankers into the current employed tanker's WPT node assignments. In addition to reducing any infeasibility, the insertion point is strongly influenced by the requirement that some RGs must be escorted over open water. For the generation of TKI moves, the tankers are placed in a "pool" for each beddown base. If a beddown base has unassigned tankers in its pool, then one of the unassigned tankers is selected. For each selected tanker, moves that insert the selected tanker into the current solution are generated. Available insertion points start before the second waypoint node assigned to any tanker and continue up to before that tanker's last waypoint node. An example of a TKI move, given the above initial solution, is "Insert tanker 5 in front of node 25". Using permutation post-multiplication, this move is represented by (0,5,25).

Performing the multiplication, (0,15,16,…,24,25,26,…,46) * (0,5,25), yields (0,15,16,…,24) (5,25,26,…, 46). Now tanker 0 is assigned nodes 15 through 24 and tanker 5 is assigned nodes 25 through 46.

Placement of the remaining tankers continues until there are no available tankers or until a feasible solution is obtained. For example, if there were 14 required escort arcs between the 32 WPTs, the 15 tankers might be assigned to the WPTs yielding the following initial solution:

(0,15)(1,16,17)(2,18,19)(3,20,21)(4,22,23,24)(5,25,26)(6,27,28)(7,43,44)(8,29,30)(9,45,46)
(10,31,32,33,34) (11,35,36)(12,37,38)(13,39,40,41,42)

where 14 tankers are used before TKI moves are no longer used.

For every iteration where the TKI move neighborhood is used, a TKI move will be generated for each WPT node and tanker base with available tankers combination for a total of |WPT nodes|*|tanker bases| moves.

The choice of the above approach was motivated by two benefits: (1) the approach is easy to implement and (2) the associated move evaluation process drives the initial solution towards excellent solutions while feasibility is simultaneously being achieved.

Once the initial TKI moves have performed their function, additional move types are invoked based on the current search status and solution. The move types discussed below were determined by both historical successes with the VRP (Carlton, 1995) and by the need to develop new move types that would exploit the specific structure of the AFRP. These move neighborhoods include: (1) return to base inserts (RTBI), (2) restricted inserts (RI), (3) escort pair inserts (EPI), (4) return to base deletes (RTBD), (5) tanker swaps (TKS), (6) restricted swaps (RS), and (7) return to base swaps (RTBS).

Each move neighborhood creates a set of eligible moves to be evaluated and all use permutation post-multiplication. Post-multiplication allows the search to investigate AFRP solutions from different conjugacy classes in S_n. In the following sections, each move neighborhood cited earlier is described with the conditions that cause them to be invoked.

6.8.1 Return to Base Insert Moves (RTBI)

After the TKI strategy is complete, the solution may still be infeasible, primarily due to a lack of tanker capacity. After the initial solution is achieved, the RTBI neighborhood is invoked whenever the current solution is infeasible. An RTBI move can be used to reduce infeasible fuel shortage at the cost of delaying one or more RGs, i.e. the tanker is unavailable until it has been refueled and allowed to reenter service.

The RTBI neighborhood is implemented by making available an RTB node for each active tanker base in the deployment scenario. These RTB nodes may be inserted within each of the current tankers' assignments. Available insertion points start at the first waypoint node assigned to a tanker within the solution and continue until after the last waypoint node. Suppose an RTB node is placed in tanker 4's assignment before node 24 for the solution

(0,15)(1,16,17)(2,18,19)(3,20,21)(4,22,23,24)(5,25,26)(6,27,28)(7,43,44)(8,29,30)
(9,45,46)(10,31,32,33,34)(11,35,36)(12,37,38)(13,39,40,41,42)

The move that will accomplish this is (24,49). The only affected cycle in this solution contains tanker 4, so only changes in that cycle need to be shown. (This convention of showing only the affected cycles will be followed throughout the rest of this chapter.) Performing this move yields (4,22,23,24) * (24,49) = (4,22,23,49,24). Tanker 4 is now allowed to return to base, after servicing nodes 22 and 23, for refueling and ancillary operations before completing its assignment by servicing node 24.

For every iteration where the RTBI move neighborhood is used, a RTBI move will generated for each node (WPTs and tankers) part of the current solution and tanker base combination for a total of |WPT nodes + active tanker nodes| * |tanker bases| moves.

6.8.2 Restricted Insert Moves (RI)

Following initial solution construction, the RI neighborhood is invoked at each iteration of the search and allows an individual node to be inserted in a different position in its cycle or to be inserted in another tanker's cycle. An RI move can either reorder a tanker's assignment or change the partitioning of the letters among the tankers. The "restriction" on this type of move limits the allowable "distance" that a letter can be moved within the current permutation solution representation. "Distance" is defined as the number of positions a letter may move from its current position either to the left or right. For the results presented in this paper, the distance was set at five positions. This parameter setting and similar parameter settings discussed later were found through empirical experimentation for the example problems studied.

Consider the following two RI example moves. First we insert a node within its current cycle, i.e. "Insert node 24 in front of node 22 within tanker 4's current assignment". The move and changes in tanker 4's assignment are achieved by (4,22,23,24) * (4,22,24) = (4,24,22,23).

Next we insert a node into another tanker's cycle., i.e. "Insert node 24 from tanker 4's assignment in front of node 25 in tanker 5's assignment". The move and changes in the tankers' assignments are: (4,22,23,24)(5,25,26) * (4,25,24) = (4,22,23)(5,24,25,26).

For every iteration where the RI move neighborhood is used, a RI move will be generated for every node (WPTs and RTBs) and position combination within five places, left or right, of the node. This will produce a total of |WPT nodes + RTB nodes| * (distance * 2) moves.

6.8.3 Escort Pair Insert Moves (EPI)

After the initial solution construction, the EPI neighborhood is invoked for each iteration of the search. The EPI neighborhood inserts the two nodes associated with an escort arc. (An escort arc connects the last node of an earlier WPT to the first node of the adjacent WPT later in the path of a RG requiring escort. In any feasible solution, this pair of nodes must be adjacent and in the correct temporal order.)

Consider the following two clarifying examples: First, we "insert the pair (31,32) after node 34 in tanker 10's assignment". The move and changes in tanker 10's assignment are (10,31,32,33,34)*(10,31,33) = (10,33,34,31,32).

Second, we "insert the pair (31,32) after node 46 in tanker 9's assignment". The move and changes in the tankers' assignments are:

(9,45,46)(10,31,32,33,34) * (9,31,33) = (9,45,46,31,32)(10,33,34)

This reassigns the escort arc duty from tanker 10 to tanker 9.

For every iteration where the EPI move neighborhood is used, an EPI will be generated for each escort pair and position combination within 5 places, left or right, of the node. This will produce a total of |escort pairs| * (distance * 2) moves.

Both the RI and EPI neighborhoods can be viewed as variants of the traditional *k*-OrOpt move.

6.8.4 Return to Base Delete Moves (RTBD)

RTBI moves help the search progress towards feasibility while increasing the number of nodes being used in the solution. Once the number of letters reaches 1.5 times the original number of nodes, the RTBD move neighborhood is invoked. This neighborhood is used to remove any extra, i.e. nonbeneficial, RTB nodes from the solution.

An illustrative example of an RTBD move is "Remove the return to base node 49 from tanker 4's assignment".

The move and change in tanker 4's assignment is: (4,22,23,49,24) * (24,49) = (4,22,23,24).

For every iteration where the RTBD move neighborhood is used, a RTBD move will be generated for each RTB node in the current solution for a total of |active RTB nodes| moves.

6.8.5 Tanker Swap Moves (TKS)

In addition to the RTBD neighborhood, the TKS neighborhood is invoked when the number of nodes in the solution has grown to 1.5 times the original number. This neighborhood allows idle tankers to be exchanged with active tankers (from different bases) to reduce travel and fuel usage. For each beddown base that has unassigned tankers in its tanker pool, a move that exchanges an assigned tanker from a different beddown base within the current solution is generated.

An example of a TKS move is "swap tanker 14 for tanker 10". The move and changes in the solution are (10,31,32,33,34) * (10,14,31) = (14,31,32,33,34).

Additionally, the TKS neighborhood allows an idle tanker to be exchanged with an active tanker that has been relocated to the idle tanker's beddown base. This may occur in two ways: (1) a tanker may relocate, provide service and then return to an active tanker base (which differs from the relocation base and beddown bases) and (2) the tanker relocates, provides service, and then returns to the relocation base.

An example of the first case is "swap idle tanker 10 for active tanker 7". Tanker 7 has relocated to tanker 10's beddown base prior to servicing WPTs. After servicing nodes 36 and 37, tanker 7 relocates to another active tanker base. If the original assignment for tanker 7 is (7,67,35,36,65) (idle tanker 10's assignment is (10)), the move and change is (7,67,35,36,65) * (7,10,35,67) = (10,35,36,65). Note that RTB node 67 also has been removed from activity.

An example of the second case is "swap idle tanker 5 for active tanker 13". Tanker 13 has relocated to tanker 5's beddown base, provided service, and then returned to tanker 5's beddown base. Suppose the original assignment for tanker 13 is (13,51,41,42,57). The move and changes are (13,51,41,42,57) * (5,41,51,13,57) = (5,41,42) and two RTB nodes, 51 and 57, have been removed.

For every iteration where the TKS move neighborhood is used, a TKS move will be generated for each tanker in the current solution and tanker base combination for a total of |active tankers|*|tanker bases| moves. Additionally, a TKS move will be generated for each active relocated tanker and tanker base combination for a total of |relocated active tankers| * |tanker bases| moves. In total, |active tankers + relocated active tankers| * |tanker bases| moves are generated.

6.8.6 Restricted Swap Moves (RS)

During the search, if a specified number of iterations have passed (20 iterations for this presentation) without a new best solution being identified, the RS neighborhood is invoked. RS moves allow an individual node within a tanker's assignment to be swapped either with a node in its cycle or with a node in another tanker's cycle. This maintains the current cardinality of the partitioning of the letters amongst the tankers. The "restriction" of this neighborhood limits the allowable distance (five positions) that any letter can be moved.

An example of an RS move where we swap a letter with another in its current cycle is "swap node 31 for node 34 in tanker 10's assignment". The move and changes of tanker 10's assignment is (10,31,32,33,34) * (10,32)(31,34) = (10,34,32,33,31).

An example of swapping one tanker's node with another tanker's node is "swap node 34 from tanker 10's assignment with node 24 of tanker 4's assignment". This yields

(4,22,23,24)(10,31,32,33,34) * (4,10)(24,34) = (4,22,23,34)(10,31,32,33,24).

For every iteration where the RS move neighborhood is used, a RS move will be generated for each node (WPTs and RTBs) and position combination within five places, left or right, of the node. This will produce a total of |WPT nodes + RTB nodes| * (distance * 2) moves.

6.8.7 Return to Base Swap Moves (RTBS)

Like the RS neighborhood, the RTBS neighborhood is invoked when a specified number of iterations (20 for this chapter) have passed without a new best solution being identified. This neighborhood allows the RTB nodes to be exchanged with other RTB nodes. This allows the solution to adjust the locations of the return to base nodes to fit the current set of tanker assignments. For each beddown base, a RTB node is available. For each selected RTB node, a move that exchanges an RTB node from a different beddown base within the current solution is generated.

An example of an RTBS move is "swap node 48 for node 56 in tanker 13's assignment". If tanker 13's assignment was (13,34,48), the move and change in tanker 13's assignment is (13,34,48) * (13,48,56) = (13,34,56).

For every iteration where the RTBS move neighborhood is used, a RTBS move will be generated for each RTB node and tanker base combination for a total of |active RTB nodes|*|tanker bases|.

AFRP-GTTS move evaluations determine, first and most importantly, whether a move yields a feasible solution and, second, whether a move yields a superior solution. Part of the power of TS is achieved by using the solution and/or move history. Wiley (2001) and Barnes et al. (2004) present in detail the search methodology implemented in AFRP-GTTS. The AFRP-GTTS memory structure stores attributes of the moves that have been previously selected and changes the length of the "short term memory" (tabu tenure) using an adaptive procedure. Move attributes are stored in a square matrix of dimension n, the number of solution nodes. As moves are implemented, the letters of the move are recorded in the matrix. For each letter of the move, the row index is the letter, the column index is the image of the letter (the "next" letter in the cyclic context of the move), and the value placed in the matrix is the current iteration plus the current tabu tenure.

At each iteration of AFRP-GTTS, move based neighborhoods are dynamically generated in accordance with the current incumbent solution and current status of the TS. Each neighboring solution requires that the associated timings of all aircraft be precisely computed for appropriate evaluation of that solution's objective function.

AMC provided a typical middle east deployment scenario to provide information on AFRP-GTTS performance on an AFRP of practical size. The best solution, found in 2 hours and 16 mins, employed 95 tankers flying 326,968 miles and allowed the latest RG to arrive in 55 hours. The tanker assignments in this solution were obtained at the *detailed asset operational level* (individual tankers or receivers) and yield excellent assignments in terms of such metrics as minimizing tanker usage, tanker travel, and total deployment time. In general, depending on the desire of the user, Wiley's AFRP method can provide either a single "best" solution to a particular problem or a set of robust, comparable solutions to a problem instance.

The fact that the best previous methodology, CMARPS, a large scale digital simulation model, requires several analysts weeks and occasionally months to achieve a single feasible inferior solution strongly underscores the power and flexibility that AFRP-GTTS adds to the "toolbox" available to AMC planners.

6.9 Air Crew Scheduling

One of the most important aspects of running an operational Air Force flying unit or major airline is scheduling flight crews. Once the Air Tasking Order (ATO) or airline schedule is published to specify the flights to be flown daily, crews must be intelligently assigned to each flight. For today's airlines, crew costs are the second highest component of direct operating costs, with the fuel cost being the highest (Gershkoff 1989). Crew costs such as temporary duty (TDY) per diem are also a significant portion of the direct operating costs of any flying unit within the USAF, but the mission of the Air Force encompasses more important concerns. Among these concerns are that many flying units operate with an insufficient number of crews. Thus, improper crew scheduling limits combat operations. When lives and national interests are on the line, any increase in the probability of mission failure is unacceptable.

In addition to cost, many other factors must be considered when dealing with the tanker crew scheduling problem (TCSP). The USAF dictates the handling of its crews through various regulations. For example, a crew's flight duty period starts when an aircrew reports for a mission, briefing, or other official duty and ends with an outbriefing following engine shut down at the completion of a mission, mission leg, or a series of missions (AFI 11-202V3, 1998). Air Force Instruction 11-202 Volume 3 (AFI 11-202V3, 1998) states that the maximum flight duty period for a tanker aircrew is 16 hours. Many other such constraints apply to the TCSP, and they are defined in the methodology section. For commercial airlines, the Federal Aviation Administration (FAA) has also established complex rules to ensure that crewmembers fulfill their duties at an appropriate performance level. These regulations, combined with the nature of the underlying combinatorial problem, make the crew scheduling problem (CSP) very difficult to solve. Gershkoff (1989) describes the airline CSP as follows:

1. The objective is to minimize the cost of flying the published schedule, subject to constraints 2–5 below.
2. Each flight must be covered once and only once.
3. Each pairing (pairings are sequences of flights a crew flies) must begin at a crew base, fly around the system, and return to the same base.
4. Each pairing must conform to the limitations of FAA regulations and published work rules in force at the airline.
5. The total number of hours flown by crews at each crew home base must be within specific minimum–maximum limits, in accordance with the airline's manpower plan.

Gershkoff details the components of crew scheduling cost for American Airlines. The USAF incurs many of the same costs, such as the hotel and per-diem expenses resulting from scheduling layovers away from each crew's home base. Gershkoff (1989) also coins a term called pay and credit, which represents unproductive crew time that must be minimized, i.e. paying crews while they are on the ground. Given the Air Force's number of tanker crews, unproductive crew time is an item Air Force operational units must minimize as well.

The formulation of the mathematical model of the CSP is based on the airline's published flight schedule, which is equivalent to the USAF ATO. The published schedule includes departure/arrival locations and times for each flight segment during a month. Flight segments are nonstop flights between pairs of cities. For tanker refueling, these flight segments consist of nonstop flights between pairs of operational bases. Refueling waypoints exist between tanker departures and arrivals, but these mid-flight stops simply add to the time length of the flight segment and do not require explicit modeling.

Constraint (2) from the airline CSP described earlier leads to formulation of the classic set partitioning problem (SPP). In a set partitioning problem, each member of a given set, S1, must be assigned to or partitioned by a member of a second set, S2. For the air crew scheduling problem, each member of the set of flights must be assigned to a member of the set of crew rotations.

The solution structure of the TCSP provides a natural partitioning of the flights. Each flight is placed in exactly one crew rotation, which represents a crew and the flights they fly in a given period of time. These disjoint rotations have a one-to-one correspondence with the columns of the set partitioning problem's constraint matrix, as seen in Figure 6.9. The disjoint rotations also represent a partial solution to the TCSP, i.e. (0,4,6,9) in Figure 6.9 is one crew rotation within the solution set of crew rotations. Throughout the TS process, these types of partial solutions can be recorded in a pool of columns for a SPP optimizer.

Once a set of columns or crew rotations is generated, the mathematical program for the SPP is as follows (Hoffman and Padberg 1993):

6.9.1 Set Partitioning Problem (SPP) Formulation

$$\min \sum_{j=1}^{n} c_j x_j$$

$$\text{subject to: } Ax = e_m,$$

$$x_j \in \{0,1\} \text{ for } j = 1, \ldots, n,$$

where e_m is a vector of m ones, and n is the number of rotations that we consider. Each row of the $mxnA$ matrix represents a flight segment, while each column represents a crew rotation with cost c_j for using it. The x_j are zero-one variables associated with each rotation, i.e. $x_j = 1$ if rotation j is flown. The A matrix is generated one column at a time, with $a_{ij} = 1$ if flight leg i is covered by rotation j, 0 otherwise.

Although the set partitioning formulation is most often used for the airline CSP, researchers have developed other formulations as well. Many airlines relax constraint (2) above and allow deadheading, typically for intercontinental flying schedules. Deadheading occurs when crews are allowed to fly on a flight segment as passengers, repositioning them for better utilization later. Graves et al. (1993) slightly change the formulation by modeling the problem as an elastic embedded SPP, allowing a flight segment to be uncovered but penalizing the solution if this constraint violation occurs. Finally, Desaulniers et al. (1997) take an entirely different approach by modeling the CSP as an integer, nonlinear, multi-commodity network flow problem (Fredley 2001).

(0,4,6,9)(1,5)(2,7)(3,8,10)

Crew 0	1	0	0	0
Crew 1	0	1	0	0
Crew 2	0	0	1	0
Crew 3	0	0	0	1
Flight 4	1	0	0	0
Flight 5	0	1	0	0
Flight 6	1	0	0	0
Flight 7	0	0	1	0
Flight 8	0	0	0	1
Flight 9	1	0	0	0
Flight 10	0	0	0	1

FIGURE 6.9 Mapping a TCSP to the SPP.

The SPP defined above is a NP-complete problem (Housos and Elmroth 1997). For as few as 1000 flight segments, billions of feasible rotations exist. Problems of this size are impossible to exhaustively enumerate and solve optimally, and this has led researchers to propose a variety of solution algorithms. These algorithms can be grouped into four categories: heuristics, optimization methods requiring a priori generation of the SPP columns, column generation approaches, and miscellaneous “other” approaches.

Interestingly, no TS approach to the airline CSP exists in the literature today. This is a noticeable absence given its success on other COPs (Dowsland 1998; Lourenco et al. 1998; Shen and Kwan 2000; Wiley 2001).

Vocabulary building is an advanced concept in the area of TS. Glover and Laguna (1997) define vocabulary building as, "Identifying meaningful fragments of solutions, rather than focusing solely on full vectors, as a basis for generating combinations." They further state, "In some settings these fragments can be integrated into full solutions by means of optimization models."

Rochat and Taillard (1995) and Kelly and Xu (1998) successfully implemented an optimization-based type of vocabulary building as they implemented different heuristic approaches to the VRP. Rochat and Taillard found augmenting their initial heuristic approach with a post-optimization set partitioning problem (SPP) solved with CPLEX MIP (ILOG, 2002) allowed them to match the best known results of many benchmark VRPs.

Kelly and Xu (1998) experienced this type of improvement as well, but they found the CPLEX MIP ran out of memory and failed to find solutions for many of their larger problems. They developed a two-phased approach to the VRP to overcome this limitation. In phase 1, they used various heuristics to develop the columns of a SPP. These heuristics typically found, at a minimum, a feasible solution to the problem. Phase 2 entailed using a TS routine they developed to solve the large partitioning problems created by phase 1.

Interestingly, both groups used their vocabulary building mechanism as a post-optimization scheme rather than embedding it into their heuristic search. Kelly and Xu (1998) suggest that finding a mechanism to integrate the column generation and SPP solution phases is "an interesting avenue of research."

Table 6.10 below describes the four main regulatory crew constraints for this problem. Crew rest is simply the minimum amount of time a crew needs to be inactive between duty periods. The 30 and 90 day flying limits represent the maximum number of hours a crew can fly during those time periods. Since crews enter an operation with a flying history, these histories must be considered when creating a current crew schedule.

Allowing crews to deboard one aircraft and take off with another creates an additional constraint within our TCSP. Clearly, a minimum time is needed for crews to leave one aircraft and operate another. Even if crews land and take off with the same aircraft, there exists some minimum time to taxi along the runway between flights. This is modeled by adding a minimum waiting time between flights (*MWBF*) constraint to the TCSP. Since no *MWBF* exists in USAF regulations, these values must be defined by the tanker crew analyst/scheduler using the approach described in this chapter.

The final constraint added to the model involves simple geography. If a crew arrives at base A, it must also depart from base A. The resulting TCSP can be described as follows:

1. Minimize the number of crews required and maximize the efficiency of the crews, subject to constraints 2–7 below. To maximize crew efficiency, the hybrid approach minimizes the time crews spend waiting on the ground.
2. Each flight of the aerial refueling problem must be flown uniquely.

TABLE 6.10 Crew Constraints for the Tanker CSP

Constraint	Limit (hours)
Flight duty day	16
Crew rest	12 (min)
30 day flying limit	125 (max)
90 day flying limit	330 (max)

3. Crew duty days must not exceed 16 hours.
4. Once their duty day is over, a crew must rest for a minimum of 12 hours.
5. Crews can fly no more than 125 hours in 30 days and 330 hours in 90 days.
6. The user-defined *MWBF* must be met.
7. Bases of arrival and departure must match for each crew and aircraft.

Combs (2002) and Combs and Moore (2004) extend the heuristic/post-optimization approaches discussed previously by developing an integrated TS/SPP optimizer and describe the SPP portion of the optimizer and how the SPP optimization is embedded within the overall TS scheme.

Our hybrid approach collects partial solutions. To initialize the pool, all the feasible or near feasible crew rotations from the heuristically generated initial solution are added to the pool.

A near feasible crew rotation may violate, within some acceptable tolerance, constraints (3) through (6) of the TCSP as defined above, with one caveat. It may not allow the *MWBF* from (6) to become negative, i.e. the crew departs on a flight before it arrived from a previous flight. The acceptable tolerance for each violable constraint is a decision-maker/analyst defined input to the solver, and represents the risk that he/she is willing to take for any particular constraint. For example, the decision maker may be comfortable allowing crew duty days to increase by 2 hours and/or crew rest to decrease by 1 hour during the scheduling period. The near feasible crew rotations are used to create near feasible solutions to the TCSP, which are ultimately presented to the decision maker/analyst as alternatives to the best feasible solutions.

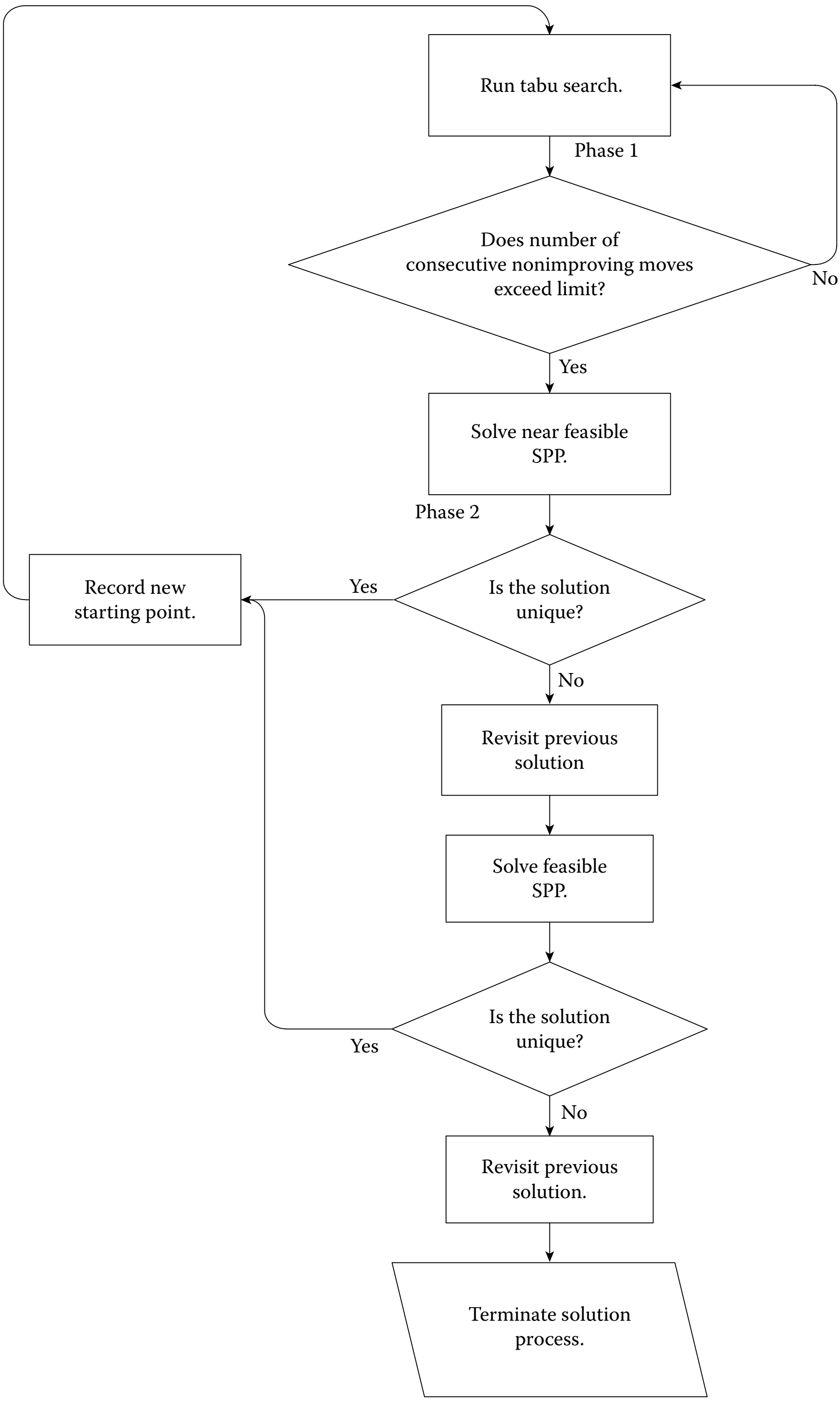
Once the pool is initialized with the individual crew rotations from the initial solution, crew rotations may be added with each neighborhood move evaluation. Since the neighborhoods created within our TS consist of swaps and inserts, only two crews are affected by any move. If the crew rotations of either affected crew are feasible or near feasible partial solutions, and the partial solutions have not been previously recorded, the pool is updated.

At various times during the search, the hybrid approach solves a near feasible or feasible SPP using the Java Concert Technology embedded within ILOG CPLEX 7.5. The near feasible SPP contains crew rotations that are both feasible and near feasible. The feasible SPP has columns whose crew rotations are all feasible. This ensures the solution created by solving a feasible SPP is itself feasible. The SPP is defined above. For the TCSP, the waiting time of each rotation represents its cost coefficient. Based on extensive testing, there are two reasons for using waiting time despite its minimization being the methodology's secondary objective:

1. Solving a SPP that minimizes the number of active crews causes the methodology to converge prematurely to local optimal solutions.
2. Feasible solutions often occur in the vicinity of near feasible solutions. Minimizing waiting times moves the search to near feasible solutions with increasingly fewer active crews. The TS itself finds smaller-crewed feasible solutions near these infeasible solutions. The hybrid approach uses the SPP optimizer to vocabulary build. The SPP's main role is to provide the search with excellent points at which to restart the search.

Finally, note that many of the SPP problems the search creates are too large for CPLEX to efficiently solve. In some cases, the solution time for CPLEX either overwhelms the overall search time or CPLEX runs out of memory and fails to report a feasible answer. Therefore, the hybrid approach actually uses CPLEX in a heuristic manner. It places a 10-minute threshold on the SPP solution process. If an optimal solution is not found in 10 mins, CPLEX reports the best solution found after 10 mins.

Finally, the rules used to invoke the SPP optimizer must be discussed. Figure 6.10 displays a flow chart of the overall process. Phase 1 of the hybrid approach simply involves running the basic TS. Once the TS reaches a preset limit of nonimproving iterations, defined by not finding a new feasible or near feasible solution, the hybrid approach starts phase 2 by calling the CPLEX solver. A near feasible SPP is always solved first. This keeps the search moving towards smaller-crewed solution space regions. If the



Note: User may terminate process at any time

FIGURE 6.10 Flowchart of the Simulation Model.

solution found is unique, a new phase one search starts from the near feasible solution produced, and the solver is recalled once a number of nonimproving iterations are completed. If the near feasible SPP solution is a revisit, then a feasible SPP is solved. If the solution to this problem is unique, the search restarts from the feasible solution. The search continues until (1) the near feasible and feasible SPP problems both find previously visited solutions or (2) the user manually terminates the process. the TS alone performs as well as the hybrid approach in terms of the number of crews and total waiting time found in the final solution. Both corresponding p-values are greater than 0.05, hence the H_0 described above is not rejected for either of those responses. But vocabulary building clearly improves the solution speed of the search process. The number of iterations to termination is reduced an average of 8459 when switching from the TS alone to the hybrid approach. The overall average masks the fact that there was a large variation in the number of iterations required to solve particular test problems. Likewise, while CPLEX took a significant amount of time to solve some of the vocabulary building SPPs, the strategy worked so well that it overcame this limitation and switching still reduced the total solution time an average of over 28 mins. This is very significant considering the average total solution time over all experimental runs is approximately 36 mins. The vocabulary building scheme creates this improvement by avoiding large portions of the solution space traversed by the TS alone.

The heuristic used to create starting solutions for the search appears to create good initial solutions. It provides solutions, on average, within 7.71% of the crew bound and within 13.51% of the waiting time bound. The TS (alone) and hybrid approach do an excellent job of improving these solutions. With respective percentage distances of 3.78 and 5.03, they reduce the distance from the crew bound by more than half and reduce the distance from the waiting time bound by nearly 63%. In addition to reducing the mean *% distance*, the solutions are much less variable as well since the standard deviation in each TCSP objective is nearly halved. This information is presented in Table 6.11.

There were many instances where the solutions found matched the lower bounds on number of crews and total waiting time. Since all solutions found were feasible, and the lower bounds show that no smaller crew or waiting time solution exists, these solutions are optimal. Indeed, the hybrid approach found the optimal solution for nearly 40% of the problems solved in the designed experiment.

The lower bounding scheme does not guarantee that the bound found is a feasible solution. Many of the hybrid approach's solutions appeared to be optimal, but they did not match the lower bounds. Therefore, the 40% optimality described above is a LB on the percentage of problems for which the hybrid approach found the optimal solution. To study this, the solution space of some of the smaller problems was completely enumerated. This showed an additional 20% of the solutions were optimal. Therefore, a total of 60% of the solutions were shown to be optimal. Of course, the TS is a metaheuristic and is not guaranteed to find the optimal solution. Only four solutions out of the 44 completely enumerated problems were shown to be suboptimal. But the results show the hybrid approach still found excellent solutions. The average *% distance* from the optimal number of crews and optimal waiting time was 0.86% and 1.00%, respectively.

While the hybrid approach performed well for the problems found in the designed experiment, a natural follow-on question is, How will the hybrid approach perform on a problem found outside the design space or a problem of significantly larger size?

The hybrid approach was applied to a large TCSP. Large is a relative term, but we label this problem as such because the schedule was over 11.5 times the size of the schedule produced by Wiley's largest, "typical" Middle-East deployment scenario (Wiley 2001). As another comparison, the TCSP contained 50% more flights than the largest benchmark airline crew scheduling problem (Beasley 2005).

TABLE 6.11 Average and Standard Deviation (SD) for % Distance with Respect to the Lower Bounds

	HA Num Crews	Starting Num Crews	HA Wait Time	Starting Wait Time
Average % distance (%)	3.78	7.71	5.03	13.51
SD % distance (%)	5.82	10.58	7.27	14.35

The problem simulates a 30-day deployment scenario. The deployment surges in the first week, generating 40 rotations each day. The tempo slows for the remainder of the schedule, generating only 15 rotations per day for the last 3 weeks. The resulting schedule contained 1269 flights, and was three times the size of the largest schedule in the experimental design.

The hybrid approach was used to solve the problem. The hybrid approach produced a feasible initial solution with 118 crews and 3,334,175 mins of total waiting time. Figure 6.11 shows the progress of the hybrid approach with respect to minimizing the number of crews and total waiting time in a feasible solution. The square boxes on the graphs represent points of vocabulary building. At the beginning of the search, little progress is made in terms of the number of crews in the solution, but total waiting time of those crews is being reduced significantly.

By iteration 365, about 27 mins after starting the solution process, a feasible solution with 3,087,760 mins of waiting time is found. This is a 7.4% reduction in waiting time, a savings of 246,415 mins or about 171 days.

At this point, the hybrid approach seems to stall and no crew or waiting time improvements are made for a long period of time. However, the hybrid approach does not terminate at any time during this stretch because the vocabulary building continues to improve the best near feasible solutions. At iteration 2421, about 3.5 hours into the solution process, vocabulary building sends the trajectory of the search into a significantly better portion of the solution space. A 117 crew feasible solution is found in the next neighborhood evaluation. The total waiting time of this solution is 2,754,589. At this point in the search, total waiting time has been reduced by 17.38%, a savings of 579,586 mins or about 402 days.

The hybrid approach continues to improve the solution until at iteration 5263, the best solution is found by solving a SPP. The best solution has 115 crews and 2,667,124 mins of total waiting time. This is a savings of three crews and a reduction in total waiting time of 20%, a savings of 667,051 mins or about 463 days! The hybrid approach terminates 255 iterations after finding the best solution. The total solution time is just over 8 hours.

6.10 Theater Distribution

The logistics distribution pipeline represents the critical link between the national industrial base and the tactical war fighter in the supported theater. This pipeline is comprised of strategic and operational components. The previous sections examined various aspects of the strategic portion of the pipeline.

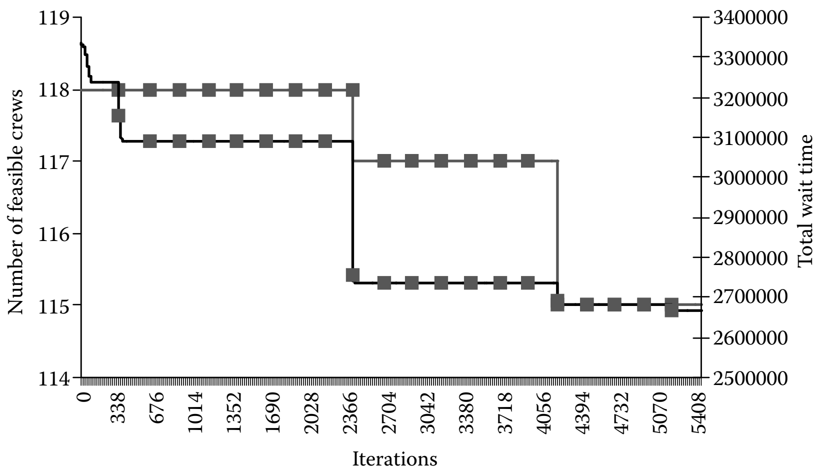


FIGURE 6.11 Progress of the hybrid approach on a large TCSP.

The theater portion of the logistics system comprises all of the networks within the theater and generally extends from the port of debarkation to the final destination at the tactical level.

Theater distribution begins at the point where material and personnel are transferred from strategic lift assets to the control of theater assets. These points are represented as APOD and sea ports of debarkation (SPOD). At the operational level of war, the senior combatant commander is responsible for development of the distribution system and ultimately for all transportation operations. Joint Publication 4-01.4 (2001) describes the theater distribution system as comprised of four networks: physical, financial, information, and communication systems. We model the physical network of the theater distribution system.

This physical network consists of the quantity, capacity, and capability of fixed structures and established facilities in support of distribution operations. This network comprises the actual infrastructure of the area of operation (AO) or theater of war and includes all roads, railroads, warehouses, supply depots, ports (air and sea), waterways, and pipelines. The network encompasses available resources such as personnel, equipment, material and the inherent ability to move these resources. The network also includes the organic distribution capability of assigned military units, commercial enterprises, host nation and multinational partners (Joint Publication 4-01.4, 2001). Figure 6.12 provides an illustrated example of a theater's physical network. This network represents a theater or area of responsibility (AOR) that has been sub-divided by the combatant commander with a theater of war. A combatant commander's AOR may contain multiple theaters of war if required by the situation. The ongoing operations in Afghanistan and Iraq represent multiple theaters of war in the CENTCOM commander's AOR.

The physical components of the network are critical to the flow of materiel and personnel in the theater. These components are categorized as nodes, edges and modes. The nodes represent the entry points of logistics into the theater of war (APOD, SPOD), the originating point of transportation assets (depots), the point where supplies are stored and processed for forward movement (transshipment) and the termination point of requirements (demands). The edges represent the "lines of communication" connecting the various nodes in the theater. The network is not necessarily fully connected and the

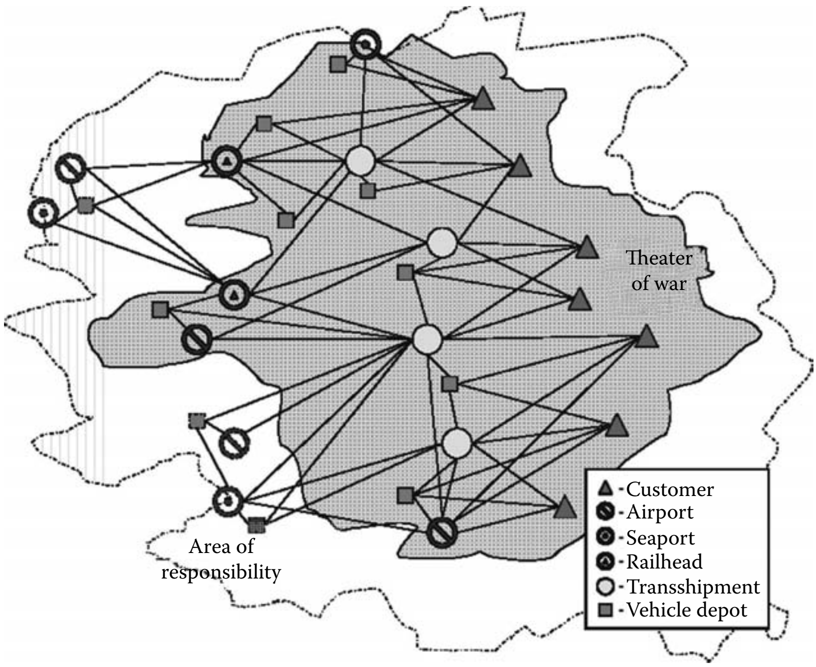


FIGURE 6.12 The theater's physical network.

connection between any two nodes is not necessarily a straight line. The theater's transportation assets are categorized as one of three mode types: air, ground, and water. Each of these modes may contain various vehicle types at the commander's disposal for distribution planning. Logistics planners utilize this collection of nodes, edges and modes to develop a distribution plan that satisfies customer demands.

This physical network possesses the basic form of a tree that is rooted at the strategic level and has its leafs at the tactical level. The number of associated levels in this tree depends on the number of logistic command levels. In general, the tree could possess as many as five layers corresponding to the rear theater facilities, forward theater facilities, and division, brigade, and battalion combat service support (CSS) units.

One way to meet the modeling requirements of the theater distribution problem is to cast it as a location, routing and scheduling pickup and delivery problem with time windows (LPDTW). This modeling approach required combining elements from both the pickup and delivery problem (PDP) and location routing problem (LRP).

Researchers have studied many versions of the PDP over the last several decades but each of these versions may be classified as one of three major PDP forms; the dial-a-ride problem (DARP), the handicapped person transportation problem (HTP), and the pickup and delivery problem with time windows (PDPTW). The general PDP is a problem of finding a set of optimal routes for a given fleet of vehicles, in order to satisfy a set of transportation requests. Each vehicle has a set cargo capacity, and a start and end location. Each request specifies the load to be transported, an origin and a destination. The set of side constraints for the PDP is more complex than those of the VRP, making it a NP-hard problem. The PDP includes a set of precedence and pairing constraints in addition to the capacity and visiting constraints in the VRP. *Precedence constraints* add the restriction that a vehicle must pickup a request before dropping it off. The *pairing*, or *coupling, constraints* further restrict the solution by requiring that the same vehicle visit both the pickup and delivery locations. Categorizing the PDP as a routing problem, scheduling problem, or routing and scheduling problem provides a clearer link to the VRP literature. A PDP scheduling problem includes both spatial and temporal factors. This additional temporal consideration converts the PDP into a pickup and delivery problem with time windows (PDPTW) (Bodin et al. 1983).

The dial-a-ride problem is an instance of the PDP that deals with transporting people instead of cargo and is the most studied class of PDP in the literature. Customers provide a request for transportation from a specific origin to a specific destination. Vehicles provide a shared service since many customers may be in a vehicle at the same time. Dial-a-ride problems generally have two conflicting objectives: minimizing operating costs and minimizing user inconvenience. User inconvenience is often measured as a deviation from a desired pick-up or delivery time or customer max riding time. We are most interested in the DARP with time windows, which is a multi-vehicle pickup and delivery problem with the added temporal concerns of time windows (m-PDPTW).

The handicapped person transportation problem is a recent line of research that is similar to the DARP except that the problem consists of different types of passengers where each type (e.g. able to walk, needs a wheelchair, etc.) requires suitable space on a vehicle from an origin to a destination (Mitrovic-Minic 1998). The service at each location occurs at a given time window and the customer's trip duration should not exceed a maximum travel time. This problem class is also an m-PDPTW but with a multi-commodity requirement based on customer type. Related research efforts include Desrosiers et al. (1986), Dumas et al. (1991), Ioachim et al. (1995), and Toth and Vigo (1997).

The location of supply facilities may be among the most critical management decisions in developing an efficient supply chain. Decisions concerning the number, size and location of these supply facilities directly impact the distribution system's cost and customer support.

Supply chain literature generally views these facility location decisions as long term strategic decisions, while routing and scheduling decisions are viewed as operational decisions (Perl and Daskin 1985). Location decisions that only consider the location of supply facilities are known as fixed charge facility location problems (FLP). The FLP seeks to determine the location of facilities and the shipping pattern between facilities and customers to minimize the total facility location and shipment costs. The location routing problem (LRP) represents an extension of the FLP.

Laporte (1988) provides a survey of early location routing procedures and provides a summary of the formulation types, solution procedures and computational results of work published prior to 1988. Min et al. (1998) is a more recent survey that provides a hierarchical and classification scheme for reviewing existing LRP literature.

Traditional research defines the LRP as a VRP with the additional requirement to determine the optimal number and locations of the depots simultaneously with solving the vehicle schedules and the distribution routes to minimize total system costs (Tuzun and Burke 1999). We are most interested in LRP solution methods that take an integrative approach to solving the LRP. The integrative approach combines a determined location planning problem with a determined route planning problem and then employs a heuristic technique to solve the combined problem. These methods are essentially location–allocation–routing (LAR) or allocation–routing–location (ARL) methods. Integrative examples from the literature include sequential, iterative and parallel methods.

Sequential methods seek to first develop a solution to either the location or routing problem and then utilize the information from this process to determine a good solution to the follow-on routing or location problem. Research for locate first and route second include Or and Pierskalla (1979), and Nambier et al. (1989). Jacobsen and Madsen (1980) provide research on the route first and locate second process.

Iterative methods fluctuate between a pure location problem and one or more route problems. The location problem addresses potential customer allocation by including approximations of delivery costs. The solution of a sub-problem provides input for another sub-problem. Perl and Daskin (1985) present several approaches that represent the iterative method.

Parallel solution methods represent the last integrative approach. The parallel solution approach attempts to nest the location and route planning phases where routes are a direct component of the location phase. Chien (1993) and Nagy and Salhi (1996) provide examples of the parallel solution method. Their approaches treat the routing problem as a sub-problem within the larger location problem.

LRP research is much more limited than either facility location or VRP research. The inherent difficulties of combining strategic or long term supply point location decisions with operational routing decisions is likely one reason that accounts for the limited attention from researchers. Wu et al. (2002) and Salhi and Fraser (1996) represent the only known works to address heterogeneous vehicles. Most literature also assumes unlimited vehicles with the same capacity. There is little research that captures the practical aspects of the LRP and no known research that captures the pickup and delivery aspects associated with the theater distribution problem (TDP).

We develop an efficient modeling approach for the TDP. The objectives of the theater distribution model change over time as the operational logistics plan transitions through the deployment, employment, sustainment and redeployment phases of the operation. However, all objectives possess a common element of determining how to best configure the distribution system to effectively and efficiently support the combat units over time. The commander's operational logistics plan establishes the physical network structure of the distribution system, provides planned locations and demands of combat forces, designates the set of all potential supply and depot unit locations, and designates the set of available transportation assets. Demands represent a multi-commodity request for delivery at a given time and transportation assets represent a fleet of heterogeneous vehicles.

The modeling objective is to determine the sequence of vehicle depot locations, allocation of transportation assets to these depots, the selection of appropriate supply points (APOD, SPOD, warehouses) and the development of a vehicle routing and scheduling plan that achieves the time definite delivery (TDD) of combat forces' demands while minimizing total system costs.

The logistics footprint of the distribution network represents the system's cost in the TDP. These costs consist of the vehicle depot and supply point establishing cost, transportation cost and necessary depot and supply point operating cost. An entity's logistics footprint (cost) is based on three factors: the weight of assigned equipment, the entity's square footage, and the number of personnel required for operations in the combat zone.

In addition to cost, the model must address many other considerations in the TDP. The establishment of a vehicle depot or supply point provides a fixed operating capacity which cannot be violated. For example, opening an airport (aircraft depot) only allows so many aircraft to operate at any given time or opening a supply point only provides a fixed (capacitated) amount of a certain commodity. Additional constraints include limiting the operation day of vehicle drivers. A driver's operating day begins with the preparation of the vehicle for operations and concludes at the end of operations maintenance. Army policy limits the vehicle operator's operational day to 10 hours. Vehicle operations beyond 10 hours require the addition of a second vehicle operator.

The problem assumes the following characteristics are known in advance: the number, location and type and quantity of demand of all customers, the potential location of all supply and vehicle depots, and the types and number of vehicles in the fleet. The TDP is further characterized as a multi-depot, multi-commodity temporal problem. The model creates a logistics distribution plan (solution) that ensures:

1. The time definite delivery of each combat force's demand is satisfied.
2. Each vehicle's route starts and ends at the same vehicle depot.
3. Each demand is serviced exactly once (multiple visits to combat force required).
4. The vehicle load at any given time never exceeds the vehicle's capacity.
5. The same vehicle picks up and delivers the demand.
6. The vehicle visits the pickup location prior to visiting the delivery location.
7. A vehicle's route does not exceed allowable travel time or distance.
8. The arrival time and departure time of any given depot location is satisfied.
9. The service time windows at the supply and demand locations are not violated.

The model simultaneously determines the number and locations of both supply points and vehicle depots, the allocation of customers to supply points and vehicles to depots, and the assignment of vehicle distribution routes and schedules, so that total system costs are minimized.

We develop an adaptive TS (ATS) heuristic for solving the TDP. A solution to an instance of the TDP is a set of vehicle routes meeting the problem's demand and timeline requirements. The model uses the cyclic form of the symmetric group, S_n, as the solution structure for the TDP. This solution structure is written as a set of disjoint cyclic factors, where each disjoint cycle represents a vehicle depot with all assigned vehicles and their associated routes and schedules. The first letter in the cycle represents the depot identification letter. Subsequent letters identify the vehicle assigned to the depot and the associated pickup and delivery points along that vehicle's route. Consider the example in Figure 6.13. This problem contains four main elements consisting of three multi-commodity demands (A, B, and C),

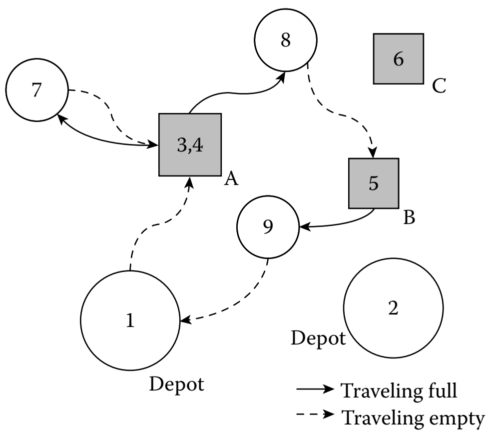


FIGURE 6.13 Graphical representation of a simple TDP.

three supply points (S1, S2, and S3), two depots (D1 and D2) and two vehicles (V1 and V2). The two potential depot locations support both vehicles but the supply points do not have the capability to support all demands. Supply point S1 can support demands A and B, while supply points S2 and S3 can only support demand C. The model's solution representation maps each of the above four elements to a letter sequentially from depots, supply points, demands and vehicles. Under this scheme, the four example elements are mapped to the following letters: depots to letters 0–1, supply points to letters 2–5, demands to letters 6–8 and vehicles to letters 9–10.

Therefore, this problem is represented using the symmetric group on 11 letters or S_{11}. Two potential solutions from S_{11} that select the depots, supply points and vehicles necessary to cover each demand are: (0,9,2,7,3,6,10,4,8), and (0,10,2,6)(1,9,3,7,5,8).

Determining the necessary network structure and associated routes requires an examination of each disjoint cycle. For example,(0,10,2,6)(1,9,3,7,5,8) represents (D1, V2, S1, A)(D2, V1, S1, B, S3, C) in the symmetric group solution and indicates that vehicle V2 is assigned to depot D1 and travels to supply point S1 to pickup the demand for delivery at location A. Also, vehicle V1 is assigned to depot D2 and travels a route between points S1, B, S3, C and back to the depot. Supply point S2 is not necessary for this operation and not included in the network distribution structure.

As the number of TDP nodes and demands increases, the size of the solution structure and problem increases. Storing full solutions or searched conjugacy class information for later retrieval becomes computationally expensive as the problem size increases. Glover and Laguna (1997) indicate that a typical solution to avoid storing all this information is the development of a hashing function that maps the desired element to an integer value. These values are stored in hash maps for utilization by the ATS search process.

Solution and move evaluation in the ATS process consists of determining a solution's feasibility and assigning it an objective function value. Solution feasibility is based on satisfying all constraints in the problem. The objective is to minimize the fixed cost, variable cost and penalty cost of the selected distribution network. Fixed cost is the total cost of establishing the set of depots and supply points plus the cost of assigning vehicles to the depots. Variable cost is the total operating cost for the selected depots and supply points plus the cost for vehicles to travel their routes. Penalty cost includes six linear violation penalties for time definite delivery (TDD) violations, demand shortfalls, route length or time violations, queue violations and storage violations. The objective function includes a solution infeasibility penalty that penalizes selecting an infeasible solution. The ATS objective function object calculates the objective function value for each solution permutation based upon the following equation, where α_i provides the ability to adjust the weight (default setting is one) of the associated factor:

$$\begin{aligned}\text{Soln}_{\text{eval}} = {} & \alpha_{\text{fc}}(\text{fixed cost}) + \alpha_{\text{vc}}(\text{variable cost}) \\ & + \alpha_{\text{TDD}}(\text{TDD penalties}) + \alpha_{\text{ds}}(\text{demand shortfall}) \\ & + \alpha_{\text{rl}}(\text{route penalties}) + \alpha_{\text{Q}}(\text{queue penalties}) \\ & + \alpha_{\text{s}}(\text{storage penalties}) + \alpha_{\text{tw}}(\text{time window penalties}) \\ & + \alpha_{\text{si}}(\text{solution infeasibility})\end{aligned}$$

Depot, supply point and vehicle assignment fixed costs represent the cost in short tons (stons) of adding the selected entity to the distribution network. Depot variable cost (stons) is an operating cost based on the number of vehicles assigned to the depot. This cost represents the overhead structure (men and equipment) necessary to manage and maintain the fleet. Supply point variable cost (stons) is an operating cost based on the supply point's daily supply throughput requirements. This represents the addition of men and equipment necessary to process increasing amounts of supplies. Vehicle variable cost (stons) is based on a vehicle's average daily route distance (km). This cost represents the second and third order support infrastructure necessary to logistically maintain the fleet of vehicles.

TDD and demand shortfall violations penalize solutions that fail to meet all customer requests. A TDD violation represents a demand not reaching its designated destination prior to the specified delivery time. The TDD violation is based on the total weight of the entire demand. There is no credit given for a partial delivery of a TDD demand requirement. Demand shortfall is the total weight in (stons) of all demands not delivered by the routing plan. TDD and demand shortfall violations are weighted according to the amount of time (mins) the demand is late. The time length for a demand shortfall is the problem's planning horizon.

Violations of route, queue, storage and time window constraints measure the feasibility of a solution. ATS considers a solution containing a vehicle route that violates one or more of these constraints a near feasible solution. A near feasible solution represents a condition or case that additional decision maker guidance may resolve. For example, the best solution found may contain a route that violates a supply point's closing time by 15 mins. This solution is near feasible and requires decision maker intervention to have the supply point remain open for an additional 15 mins. This decision renders the solution feasible. ATS tracks both the best feasible (no violations) and near feasible solutions found during the search. This provides the decision maker with an opportunity to make a near feasible solution feasible by loosening a violated condition or constraint.

Vehicle route violations are based on both route length (km) and route travel time (mins). Any given vehicle has a maximum allowed travel distance (km). This travel distance is the minimum distance of either the vehicle's on board fuel capacity or a doctrinally constrained limit. A vehicle's fuel capacity represents a hard constraint that a decision maker can overcome by allowing refueling to occur outside the vehicle's assigned depot. A travel distance violation is the difference between the actual route's distance and maximum allowed distance for the assigned vehicle and is weighted by the vehicle's capacity (stons). A given vehicle's crew possesses an operating day constraint that limits the total operating time (mins) for the vehicle in an operating day. Operating day violation is based on the difference of the time necessary to complete the route and a crew's operating day and is weighted by the assigned vehicle's capacity (stons). A route length violation represents the sum of both the travel distance and operating day violations. Queue violation is the sum of all queue violations at all nodes in a planning period. Queue violation is based on a node's ability or capacity to concurrently process vehicles at the site. This queue capacity is usually represented as a maximum on the ground or MOG for an airfield but may also represent the maximum number of vehicles a warehouse can concurrently load or offload. The queue violation is weighted by the capacity (stons) of the vehicle arriving at a node with a full queue and the length of time (mins) the node's queue capacity is exceeded. The default ATS setting forces vehicles to wait (FIFO) until a space opens at a site. Storage violation is the sum of all storage violations (stons) at all supply nodes in a planning period and the sum of all supply point throughput violations (stons). A storage violation represents a routing plan that delivers more commodities to a transshipment point than the point can actually support. This violation usually occurs in a given time period to alleviate distribution burdens in a following time period. A storage violation is the difference between the current on-hand quantity and the supply point's storage capacity for a given planning period. Storage violation also includes a penalty for violating the throughput capability of a supply point. This violation represents a supply point processing more supplies than the sites facilities or assigned personnel could actually support.

Solution infeasibility indicates that a solution permutation violates one of the following critical constraints: vehicle capacity, supply point and demand mismatch, or vehicle route mismatch. Solution infeasibility provides ATS the opportunity to temporally select an infeasible solution if it allows the search process to traverse to a better region of the solution space. These constraints represent hard infeasibilities, which are situations where the solution is meaningless based on current conditions. For example, a vehicle capacity violation indicates that a solution contains a vehicle carrying a load that exceeds its capacity. This violation represents a situation that cannot be resolved by decision maker input. A supply point and demand mismatch violation indicates that a vehicle is expected to pickup a given demand from a supply point that does not support the demanded commodity. The final infeasibility constraint represents trying to assign a vehicle to an incompatible route. For example, a C-17 aircraft cannot support a route that contains a nonairfield node.

The TS manager determines the search path for the ATS process. The strategy reviews the current parameters and determines the phase and search context for the next iteration. The manager controls movement between the request assignment (partitioning) group and the route improvement (ordering) group based on the number of iterations since the last iteration resulting in the best solution found to that point in the search. The search context is based on the current status of tracking parameters and transitions between super-diversify, diversify, and intensify. Intensification searches are focused within conjugacy classes that contain good solutions. Diversification occurs when the current search path fails to find good solutions or the process has detected an attractor basin. Phase I ends by establishing the current solution as the new incumbent solution.

The TS strategy manager determines the appropriate phase search context and move neighborhood to apply to the incumbent solution. These decisions are based on collected search data, objective function values, and pre-defined search parameters. Two move manager schemes are used. The first search scheme does *not* utilize elite lists during the intensification portion of the search process. The second search scheme maintains two elite lists to store solution data for both the best feasible and near feasible solutions found during the search process. The elite list size is a user pre-defined parameter and has a default size of five for this research. Each elite list is sequenced in descending order based on solution objective function values. ATS utilizes the elite list as a restart mechanism during the second phase of the search process. The search manager monitors the *iteration since last good solution* (ISGS) parameter that maintains the number of iterations since the search process last found a new best (feasible or near feasible) solution. The *consecutive infeasible iterations* (CI) parameter tracks the consecutive number of iterations resulting in infeasible solutions and is used by the search manager to diversify. The *choose first improving move* (CFIM) is a Boolean parameter that controls the move manager's selection process. The move manager selects the first improving move from the move neighborhood scheme when the CFIM parameter is true. This serves as a diversification scheme allowing visits to areas of the solution space the ATS process might not normally visit. When the CFIM parameter is set to false, the move manager examines all moves in the neighborhood scheme before selecting the best move.

Figure 6.14 illustrates the move manager process for establishing the search phase and search context where ATS is utilizing elite lists. The search process contains two main phases consisting of an exploration and intensification phase. Once the move manager selects the appropriate search phase, it determines the search context (intensify, diversify, or super-diversify) based on solution characteristics.

The ATS search process begins in an exploration phase and the CFIM parameter is set to true. This allows the search process to traverse a greater number of regions in the solution space. The search manager tracks both the best feasible and best near feasible solutions discovered during the search process. Each iteration, if the solution is a new best solution (feasible or near feasible), the manager updates the appropriate elite list and resets the ISGS and CI counters. The discovery of a new best solution might be an indication of a promising region so the search process enters a mini-intensification search phase to explore the local area. If the incumbent solution is not a new best solution, the search manager increments the ISGS counter and checks to see if the solution is infeasible. The move manager increments the CI parameter for an infeasible incumbent solution and resets the CI parameter if the solution is feasible or near feasible. The search manager next determines if the ISGS counter has surpassed a pre-defined tolerance or number of poor solutions.

The search manager checks the CI tolerance level if the ISGS counter has not surpassed its tolerance. Surpassing the CI tolerance indicates that the search process is exploring an undesirable region of the solution space. This triggers a solution repair phase that targets infeasibility. Once the ISGS parameter surpasses this tolerance, the search process enters a more focused search phase. The CFIM parameter is set to false allowing the move manager to examine all moves and the elite list solutions are activated.

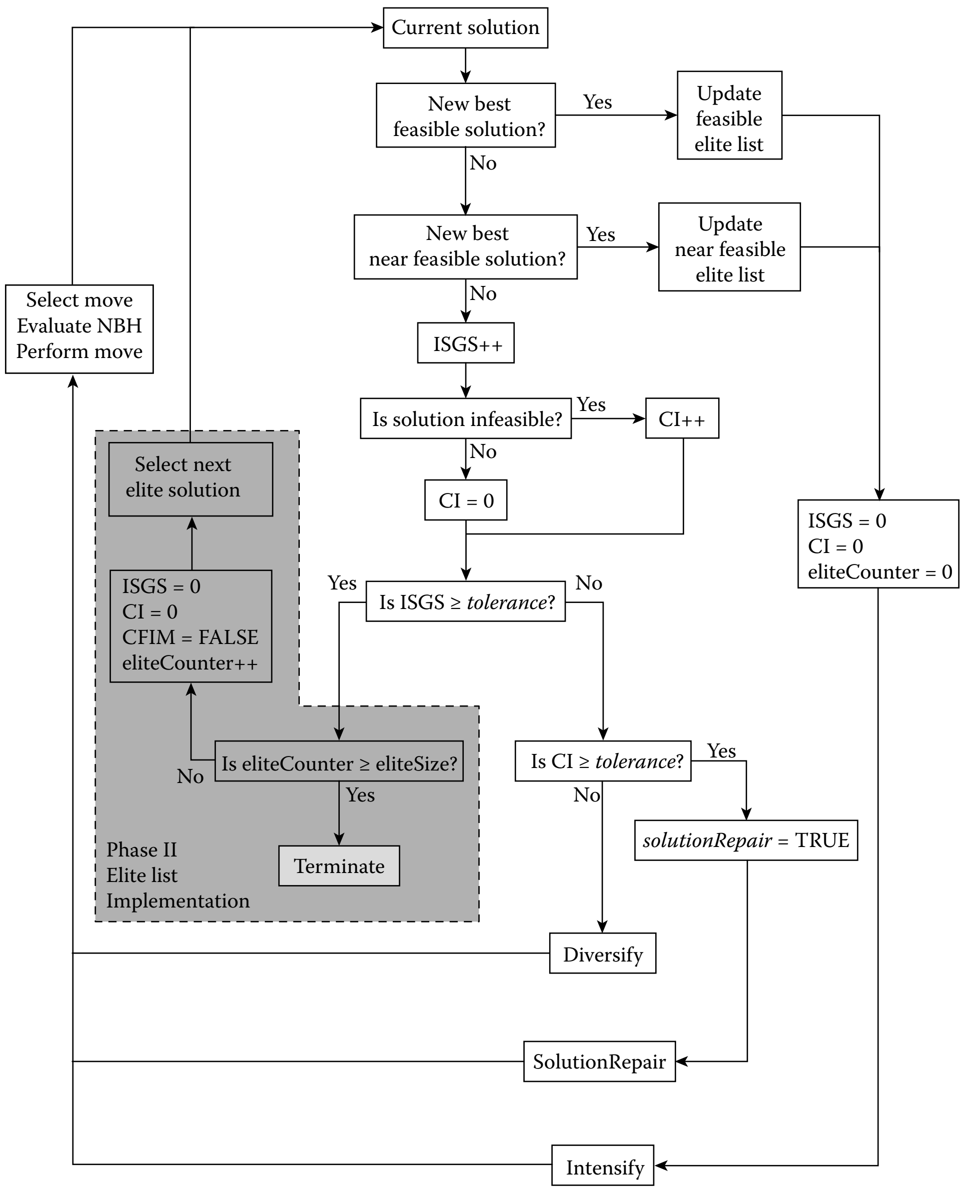


FIGURE 6.14 ATS strategy move manager.

The ATS process selects the first solution in the elite list as the new incumbent solution and repeats the phase one search process until the ISGS parameter surpasses the pre-defined tolerance. At this point, the next solution is selected from the elite list and the cycle begins again until exhausting all elite list solutions.

6.10.1 A Theater Distribution Modeling Example

ATS was applied to a set of four different theater distribution scenarios. These scenarios were based on six unit of action scenario vignettes presented in TRADOC Pamphlet 525-3-90 (2002) and represent operationally realistic problem instances for theater distribution support of the future combat system equipped unit of action. TRADOC originally designed these vignettes to test new tactical concepts and organizational design principles, and they provide a good platform to examine the ATS methodology. The intent is to investigate the efficacy of the ATS methodology and demonstrate its ability to capture the complexities of realistic problems.

The following elements capture the key problem characteristics of each scenario: the number of vehicles, number of depots, number of supply points, number of demand locations, and total number of multi-commodity demands. Table 6.12 displays the four scenarios and their key elements. The in/out for depots and supply points refers to the number of nodes located inside and outside the AO. The vehicle column refers to the total number of available ground or air transportation assets and may include several vehicle model types. For example, scenario TDP 1 includes two vehicle types based on the M871 tractor-trailer and a palletized load system ground vehicle and three aircraft types based on the Army CH47, Air Force C-130 and C-17. Scenario TDP 2 contains three versions (A, B, C) that represent the same underlying problem requirement but each of the three entries represents different distribution concepts. The intent of TDP 2(A, B, C) is to demonstrate ATS' ability to support what-if analysis and capture changes in the underlying distribution system.

TDP 2(A, B, C) provides an excellent problem set to illustrate the ATS methodology. TDP 2(A, B, C) actually represents three individual problems that differ only in the distribution policy or availability of additional distribution vehicle types.

TDP 2 represents a mid-intensity small scale contingency type operation conducted in a highly compartmentalized AO. The southern regions of Afghanistan or the areas of Trans-Caucasus provide an excellent representation of the intended difficult terrain in the AO. Logistics planners are required to determine the support structure and routing requirements necessary to sustain the units operating in the region.

Figure 6.15 provides a graphical representation of TDP 2. The problem's AO is approximately 30,000 square miles in size and currently contains one notional unit of employment (UE) that has three units of action (UAs) in the field. The AO is completely surrounded by the theater of operation and contains a single *no fly zone*. The presence of the *no fly zone* and the region's difficult terrain represent two reasons why the ATS methodology utilizes associated ground and air distance matrices versus relying on straight line distances. The problem contains 20 UA demand locations. However, C-130 type aircraft are capable of landing at only 50% of these locations. The problem also contains four potential depot locations and eight potential supply storage and distribution points. The entire distribution network is represented by a fully connected (32 nodes) ground and air graph containing 2048 arcs. This problem contains three times the number of locations and approximately five times as many arcs as Crino's (Crino et al., 2004) largest theater distribution network. Each supply point and demand location possesses a time window of operation. This 100% time window requirement places an additional burden on the routing and scheduling aspect of the problem. The problem has 100 multi-commodity (a total of six commodities) configured load demands every 24 hours uniformly distributed across the 20 demand locations. Table 6.13 provides an example set of 100 requirements (demands) for this problem.

TABLE 6.12 Scenario Problem Parameters

Scenario	Depots (In/Out)	Supply Points (In/Out)	Demand Locations	Number of Demands	Vehicles (Ground/Air)	Scenario Emphasis
TDP 1	4/0	8/0	20	100	20/60	Army vs AF Air Spt
TDP 2A	2/2	6/2	20	100	50/0	Traditional Ground Spt
TDP 2B	2/2	6/2	20	100	50/40	Limited Airfields
TDP 2C	2/2	6/2	20	100	50/40	Aerial Delivery
TDP 3	3/1	3/1	8	200	20/60	Remote Location Spt
TDP 4	0/7	2/2	7	300	0/150	UA Attack Support

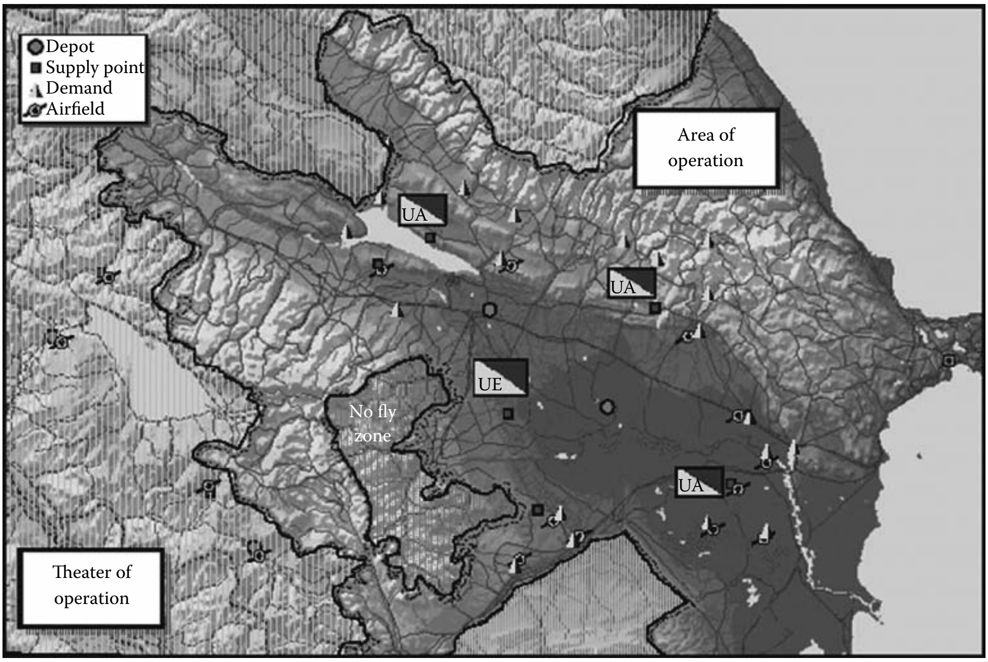


FIGURE 6.15 TDP2 graphical representation.

TABLE 6.13 Example TDP 2(A, B, C) Demand Requirements

	Demand		Delivery Window			Location Coord	
Dmd Pnt ID Ltr	Type (Class)	Quantity (STONS)	Open (MINS)	Close (MINS)	TDD Required	(x)	(y)
0	I	4.8					
	II	3.5	412	772	true	85	108
	V	10					
	IX	2					
1	I	4.8	459	819	true	37	32
	II	3.5					
	III(P)	0.5					
	V	8					
	IX	2					
2	I	4					
	II	2.5					
	III(P)	1.5	481	841	true	43	147
	V	11.5					
	IX	2.5					
3	II	2.5					
	III(P)	2.5					
	IV	4	488	848	true	8	30
	V	10					
	IX	3					

The TDP 2 problem was solved four times under the following conditions:

1. OPLOG planner algorithms with traditional hierarchal support relationships.
2. TDP 2A, ATS method with only ground distribution assets available.
3. TDP 2B, ATS method with a mix of air/ground assets available.
4. TDP 2C, ATS method with air/ground assets and the Army aerial delivery system available.

These four conditions allow for a direct comparison between the distribution system created by the current solution method (OPLOG planner) and the resulting ATS changes in the distribution system from changing the distribution concept or the addition of new equipment. Table 6.14 contains the ATS objective function values for TDP 2 under the four evaluated conditions. Condition 4, in Table 6.14 contains two entries. The first value represents the logistics footprint cost for the theater of operation and the second value is the logistics footprint cost for the AO.

Condition 1 was solved using OPLOG planner algorithms. Condition 1 represents logistics distribution under the current hierarchal support doctrine utilizing OPLOG planner algorithms to determine the distribution system. The hierarchical support concept implies that each UA possesses a dedicated supporting element and each of these supporting elements is supported by a higher supply source. The problem contained a fixed fleet of 50 Palletized Load System (PLS) type ground vehicles, with the assumption that all vehicles were available at time zero. The resulting OPLOG planner network was evaluated using the ATS objective function. The OPLOG planner distribution network opened four supply distribution points and two vehicle depots. Demand distribution required the utilization of 25 PLS type vehicles. The established network achieved all time window and time definite delivery requirements. However, the solution violated supply point constraints. The storage violation and demand shortfall penalty is the result of throughput violations at two of the four supply distribution points. These violations are an example of the planning inefficiencies associated with the current distribution methodology and OPLOG planner. OPLOG planner's inability to construct a routing and scheduling plan forces logistics planners to make manual adjustments to correct for discrepancies. Since there is a throughput violation, a logistics planner is now required to manually juggle demands by reallocating shortfalls to other distribution points. This reallocation results in a logistics footprint increase for the supply distribution points and vehicle routes that may result in violations of other constraints. Several OPLOG planner attempts, over the course of 60 mins, were made to select different supply point and route combinations but each result contained one or more ATS objective function violations.

TABLE 6.14 TDP 2 Objective Function Values

	TDP 2 Conditions			
Criterion	1	2	3	4*
Total depot fixed cost	1,000	1,000	1,000	1000/0
Total depot variable cost	250	230	110	290/0
Total supply point fixed cost	6,000	3,000	1,500	1000/0
Total supply point variable cost	6,850	3,760	3,040	4,500/0
Total vehicle fixed cost	25	23	20	29/0
Total vehicle variable cost	13,106	22,942	9,624	70,353/0
TDD violation	0	0	0	0
Demand shortfall penalty	208,800	0	0	0
Route length violation penalty	0	0	0	0
Time window violation penalty	0	0	0	0
Storage violation penalty	290,000	0	0	0
MOG violation penalty	0	0	0	0
Total objective function value	526,031	30,955	15,294	77,172/0

* Logistics footprint cost theater of operation/logistics footprint cost area of operation.

Conditions 2–4 were solved using the ATS methodology. Condition 2 (TDP 2A) utilizes the same problem characteristics as those in Condition 1 with one exception. Distribution support is changed from a hierarchal support concept to a LPDPTW support condition. The major difference is that a requestor no longer has a dedicated support element. The only requirement is the timely delivery of the requested items. ATS performed a total of 22,276 iterations. The best objective function value presented in Table 6.14 was found at iteration 1081 in approximately 20 mins. The ATS solution opened five supply points and the same two vehicle depots as Condition 1. Demand distribution required 23 PLS type vehicles but a much different routing scheme as compared to Condition 1. The selection of a different combination of supply points and a much longer routing scheme provides a feasible solution and a likely lower bound logistics footprint solution for a hierarchal distribution solution (Condition 1). The ATS approach dominated the manual planning method both in terms of solution feasibility and solution time.

Condition 3 (TDP 2B) utilizes the same problem characteristics as those in Condition 1 with two exceptions. The first is the LPDPTW support condition and the second exception is the addition of 40 C-130 aircraft to the fleet with the assumption that all aircraft are available at time zero. These exceptions allow exploring the impact of permitting air delivery on the associated distribution network. ATS performed a total of 24,296 iterations. The best objective function value, presented in Table 6.14, was found at iteration 20,745 in approximately 461 mins. However, ATS found a feasible objective function solution value of 15,649 (within 2.4% of the best solution found) in 91 mins and a feasible objective function solution value of 19,220 in approximately 24 mins. The major differences in these two solutions were the choice of delivery vehicles and their routes. ATS opened four supply points and all four vehicle depots to establish the distribution network. The major difference from Conditions 1 and 2 is that ATS opened the two depots and supply points in the theater of operation. The addition of aircraft to the problem made opening these nodes feasible. The major advantage in opening these theater nodes is the direct reduction of the required logistics footprint in the area of operation. The solution provides a direct measurement of the benefits associated with adding aircraft consideration to the planning process, an item missing from the current OPLOG planning process. The distribution of supplies utilized nine aircraft assigned to the two depots in the theater of operation and 11 PLS vehicles assigned to the two depots in the area of operation. These 11 ground vehicles were still necessary since only 50% of the demand locations possessed airfields. However, their associated logistics footprint cost is significantly better than the best logistics footprint of Condition 2 (TDP 2A). There is a clear benefit associated with allowing logistics support for the nine aircraft so the aircraft can reside outside the area of operation.

Condition 4 (TDP 2C) utilizes the same problem characteristics as Condition 3 (TDP 2B) with one exception. The problem now allows the use of the Army's experimental aerial delivery system. This system is designed to deliver air dropped cargo with precision accuracy to a requested location. This additional capability now makes it possible to support all demand locations with aircraft and determine the potential effectiveness of the new delivery system in terms of a reduced logistics footprint in the area of operation. The objective function value in Table 6.14 provides a value for both the logistics footprint in the theater of operation and the area of operation. There is no required logistics footprint in the area of operation. ATS opened both depots and the supply points in the theater of operation and utilized 29 C-130 aircraft for the distribution. This number of aircraft more than doubled the required number of aircraft of Condition 2 (TDP 2A), but the benefit lies in moving all support requirements outside the area of operation. There is still an associated logistics footprint requirement for supporting the demands but these requirements no longer concern the UE commander. For example, the vehicle variable cost for supporting the 29 C-130s is 70,353. This variable cost is much higher than that of any other evaluated Condition and assumes that the assigned depots can support the operational sustainment requirements of the C-130s.

The results for the theater distribution problems represent a significant improvement over the current methodology used in the field. Our solution technique simultaneously solves both the location of depots and the routing of vehicles for the underlining location routing problem. Traditional LRP solution methods seek to either first provide a solution to the location of the depots and then solve the corresponding routing problem (locate first, route second) or first solve the routing problem and then determine the depot locations (route first, locate second). The ATS methodology simultaneously considers both location and routing moves as it seeks to improve the overall solution cost. The ATS methodology also considers both vehicle and customer allocation to depots simultaneously with the location and routing aspects of the LRP.

The ATS methodology we developed provides an analytic foundation for determining the theater distribution network and necessary vehicle routing and scheduling requirements to ensure the time definite delivery of demands, as opposed to the current spreadsheet methods used in the field. The ATS methodology provides military analysts a tool for rapidly conducting what-if analysis for planning operations.

6.11 Conclusion

We have developed specific heuristic approaches to model particular phases of the deployment and distribution system. We have created models for various problems associated with this system: the 2D BPP, the pallet packing problem, the ACP, the SAP, the aerial tanker refueling problem, the tanker crew scheduling problem, and the theater distribution problem. In each problem area, we have achieved excellent results. The challenge is to incorporate additional real-world complexities into the models while maintaining reasonable execution times and to model other aspects of the deployment and distribution problem such as the tanker employment problem.

References

AFDD 2-6.2 1999. Air Refueling: Air Force Doctrine Document 2-6.2, Part of Joint Publication 3-17, Joint Tactics, Techniques, and Procedures for Air Mobility Operations.

AFI 11-202V3. 1998. Flying Operation, General Flight Rules, Air Force Instruction 11-202, Volume 3.

Automated Airload Planning System, http://www.tis.army.mil/AALPS/default.htm (Accessed: 2006, August 15).

Baker, B.S., E.G. Coffman Jr., and R.L. Rivest. 1980. Orthogonal packing in two dimensions. *Siam Journal on Computing* 9, 846–855.

Ballew, B.P. 2000. *The Distributor's Three-dimensional Pallet-packing Problem: A Mathematical Formulation and Heuristic Solution Approach*. Masters Thesis, AFIT/GOR/ENS/00M-02. Air Force Institute of Technology (AU), Wright-Patterson AFB OH.

Baltacioğlu, E. 2001. *Distributor's Three Dimensional Pallet Packing Optimization*. Masters Thesis, Air Force Institute of Technology, Wright-Patterson AFB OH.

Baltacioğlu, E., J.T. Moore, and R.R. Hill. 2006. The distributor's three-dimensional pallet-packing problem: A human intelligence-based heuristic approach. *International Journal of Operational Research*, 1 (3), 249–266.

Barnard, T. and H. Neill. 1996. *Mathematical Groups*. Berkshire: Cox and Wyman Ltd.

Barnes, J.W., S. Dokov, B. Dimova, and A. Solomon. 2003. The theory of elementary landscapes. *Applied Mathematics Letters*, 16 (3), 33–343.

Barnes, J.W., V.D. Wiley, J.T. Moore, and D.M. Ryer. 2004. Solving the aerial fleet refueling problem using group theoretic tabu search. *Mathematical and Computer Modelling* 39 (6–8), 617–640.

Baumslag, B. and B. Chandler. 1968. *Group Theory: Schaum's Outline Series*. McGraw Hill, New York.

Beasley, J.E. 2005. OR-Library, Container Loading. Available: http://www.brunel.ac.uk/depts/ma/research/jeb/orlib/thpackinfo.html (Accessed: 2005, February 8).

Berkey, J.O., and P.Y. Wang. 1987. Two dimensional finite bin packing algorithms. *Journal of the Operational Research Society,* 38, 423–429.

Bischoff, E.E., and W.B. Dowsland. 1982. An application of the micro to product design and distribution. *Journal of the Operational Research Society*, 33 (3), 271–281.

Bischoff, E.E., F. Janetz, and M.S.W. Ratcliff. 1995. Loading pallets with nonidentical items. *European Journal of Operational Research*, 84 (3), 681–692.

Bischoff, E.E., and M.D. Marriott. 1990. A comparative evaluation of heuristics for container loading. *European Journal of Operational Research*, 44 (2), 267–276.

Bischoff, E.E., and M.S.W. Ratcliff. 1995. Issues in the development of approaches to container loading. *Omega*, 23 (4), 377–390.

Bhatia, A.K., and S.K. Basu. 2004. Packing bins using multi-chromosomal genetic representation and better-fit heuristic. *Lecture Notes in Computer Science,* 3316, 181–186.

Bodin, L., B. Godden, A. Assad, and M. Ball. 1983. Routing and scheduling of vehicles and crews: The state of the art. *Computers and Operations Research,* 11 (2), 63–211.

Bortfeld, A., and H. Gehring. 1997. *Ein tabu search-Verfahren mit schwach heterogenem Kistenvorrat*, Technical Report 240, FB Wirtschaftswissenschaft, Fern Universität Hagen.

Burks, R.E. 2006. *An Adaptive Tabu Search Heuristic for the Location Routing Pickup and Delivery Problem with Time Windows with a Theater Distribution Application*. Ph.D. Dissertation, Air Force Institute of Technology, Wright-Patterson AFB OH.

Burks, R.E., J.T. Moore, and J.W. Barnes. 2008. Solving the theater distribution problem with tabu search. *Military Operations Research*, submitted.

Capehart, S.R. 2000. *A Tabu Search Metaheuristic for the Air Refueling Tanker Assignment Problem*. Master Thesis, Air Force Institute of Technology, Wright-Patterson AFB OH.

Carlton, W.B. 1995. *A Tabu Search to the General Vehicle Routing Problem*. Ph.D. Dissertation, The University of Texas at Austin.

Chazelle, B. 1983. The bottom-left bin-Packing heuristic: An efficient implementation. *IEEE Transactions on Computers*, c-32 (8), 697–707.

Chien, J.W. 1993. Heuristic procedures for practical sized uncapacitated location-capacitated routing problems. *Decision Sciences,* 24 (5), 995–1021.

Chocolaad, C.A. 1998. *Solving Geometric Knapsack Problems using Tabu Search Heuristics*. Masters Thesis, Air Force Institute of Technology, Wright-Patterson AFB OH.

Chu, P.C., and J.E. Beasley. 1998. A genetic algorithm for the multidimensional knapsack problem. *Journal of Heuristics*, 4(1), 63–86.

Cochard, D.D., and K.A. Yost. 1985. Improving utilization of Air Force cargo aircraft. *Interfaces* 15 (1), 53–68.

Colletti, B.W. 1999. *Group Theory and Metaheuristics.* Ph.D. Dissertation, The University of Texas at Austin.

Combs, T.E. 2002. *A Combined Adaptive Tabu Search and Set Partitioning Approach for the Crew Scheduling Problem with an Air Tanker Crew Application.* Ph.D. Dissertation, Air Force Institute of Technology, Wright-Patterson AFB OH.

Combs, T.E., and J.T. Moore. 2004. A hybrid tabu search/set partitioning approach to tanker crew scheduling. *Military Operations Research*, 9 (1), 43–56.

Crino, J. 2002. *A Group Theoretic Tabu Search Methodology for Solving the Theater Distribution Problem*. Ph.D. Dissertation, Air Force Institute of Technology, Wright-Patterson AFB OH.

Crino, J.R., J.T. Moore, J.W. Barnes, and W.P. Nanry. 2004. Solving the military theater distribution vehicle routing and scheduling problem using group theoretic tabu search. *Mathematical and Computer Modelling,* 39 (6–8), 599–616.

de Castro Silva, J.L., N.Y. Soma, and N. Maculan. 2003. A greedy search for the three-dimensional bin packing problem: the packing static stability case. *International Transactions in Operations Research*, 10 (2), 141–153.

Defense Transportation Regulation, DOD Regulation 4500.9-R-Part III (DTR 450.9-R-Part III). 2004. Mobility.

Dell'Amico, M., S. Martello, and D. Vigo. 2002. A lower bound for the nonoriented two-dimensional bin packing problem. *Discrete Applied Mathmatics,* 118, 13–24.

Dell'Amico, M., and M. Trubian. 1993. Applying tabu search to the job-shop scheduling problem. *Annals of Operations Research,* 41, 231–252.

Desaulniers, G., J. Desrosiers, Y. Dumas, S. Marc, B. Rioux, M.M. Solomon, and F. Soumis. 1997. Crew pairing at Air France. *European Journal of Operational Research*, 97, 245–259.

Desrosiers, J., Y. Dumas, and F. Soumis. 1986. A dynamic programming solution of the large-scale single vehicle dial-a-ride problem with time windows. *American Journal of Mathematical and Management Science,* 6, 301–325.

Dowsland, K. 1996. Simple tabu thresholding and the pallet loading problem. In: *Meta-Heuristics Theory and Application,* edited by I.H. Osman and J.P. Kelly. Norwell, MA; Kluwer Academic Publishers, 381–405.

Dowsland, K. 1998. Nurse scheduling with tabu search and strategic oscillation. *European Journal of Operational Research*, 106, 393–407.

Dumas, Y., J. Desrosiers, and F. Soumis. 1991. The pickup and delivery problem with time windows. *European Journal of Operational Research,* 54, 7–22.

Eglese, R.W. 1990. Simulated annealing: A tool for operational research. *European Journal of Operational Research,* 46, 271–281.

Faina, L. 2000. A global optimization algorithm for the three-dimensional packing problem. *European Journal of Operational Research*, 126 (2), 340–354.

Fraleigh, J.B. 1976. *A First Course in Abstract Algebra*. Reading MA.: Addison-Wesley.

Franks, T. 2002. Statement of General Tommy R. Franks, Commander, US Central Command. 31 July 2002. http://armed-services.senate.gov/statemnt/2002/July/Franks.pdf.

Fredley, M. 2001. A Decomposition Approach for the Multi-Modal, Resource-Constrained, Multi-Project Scheduling Problem with Generalized Precedence and Expediting Resources. Ph.D. Dissertation, Air Force Institute of Technology, Wright-Patterson AFB OH.

Garey, M.R., and D.S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco, CA: W.H. Freeman.

Gehring, H., and A. Bortfeld. 1996. *Ein genetischer Algorithmus für das Containerbe-Ladungsproblem*, Technical Report 227, FB Wirtschaftswissenschaft, Fern Universität Hagen.

Gehring, H., K. Menschner, and M. Meyer. 1990. A computer-based heuristic for packing pooled shipment containers. *European Journal of Operational Research*, 44 (2) 277–289.

Gendreau, M., G. Laporte, and G. Potvin. 1997. *Vehicle Routing: Modern Heuristics.* Chichester, UK: John Wiley & Sons, 311–336.

George, J.A., and Robinson, D. F. 1980. A heuristic for packing boxes into a container. *Computers and Operations Research*, 7 (1), 147–156.

Gershkoff, I. 1989. Optimizing flight crew schedules. *Interfaces*, 19 (4), 29–43.

Glover, F., and M. Laguna. 1997. *Tabu Search*. Boston: Kluwer Academic Publishers.

Graves, G.W., R.D. McBride, I. Gershkoff, D. Anderson, and D. Mahidhara. 1993. Flight crew scheduling. *Management Science*, 39 (6), 736–745.

Guéret, C, N., O. Jussien, C. Lhomme, C. Pavageau, and C Prins. 2003. Loading aircraft for military operations. *Journal of the Operational Research Society*, 54, 458–465.

Han, C.P., K. Knott, and P.J. Egbelu. 1989. A heuristic approach to the three-dimensional cargo-loading problem. *International Journal of Production Research*, 27 (5), 757–774.

Harder, R.W. 2000. *A Jump Search Enhanced Reactive Tabu Search in a Universal Vehicle Router Applied to Routing Unmanned Aerial Vehicles*. Master's Thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH.

Harder, R.W. 2002. OpenTS: http://www.iharder.net (Accessed: 2006, February 12).

Harder, R.W., R.R. Hill, and J.T. Moore. 2004. A Java universal vehicle router for routing unmanned aerial vehicles. *International Transactions in Operational Research*, 11 (3), 259–276.

Harwig, J.M. 2003. *An Adaptive Tabu Search Approach to Cutting and Packing Problems.* Ph.D. Dissertation, The University of Texas at Austin.

Harwig, J.M., J.W. Barnes, and J.T. Moore. 2006. An adaptive tabu search approach for 2-dimensional orthogonal packing problems. *Military Operations Research,* 11 (2), 5–26.

Hebert, A.J. 2006. Air mobility's never-ending surge. *Air Force Magazine,* September, 46–52.

Heidelberg, K.R., G.S. Parnell, and J.E. Ames, IV. 1998. Automated air load planning. *Naval Research Logistics,* 45, 751–768.

Hill, R.R., Yong Kun Cho, and J.T. Moore. 2008. New problem reduction heuristics for the multidimensional knapsack problems based on the core problem. *European Journal of Operational Research*, submitted.

Hoffman, K.L., and M. Padberg. 1993. Solving Airline Crew Scheduling Problems by Branch-and-Cut. *Management Science*, 39(6), 657–682.

Housos, E., and T. Elmroth. 1997. Automatic optimization of subproblems in scheduling airline crews. *Interfaces*, 27 (5), 68–77.

ILOG. 2002. http://www.ilog.com (Accessed: 2002, May 23).

Ioachim, I., J. Desrosiers, Y. Dumas, and M. Solomon. 1995. A request clustering algorithm for door-to-door handicapped transportation. *Transportation Science,* 29(63), 78.

Isaacs, I.M. 1994. *Algebra*. Pacific Grove, CA: Brooks/Cole Publishing.

Jacobsen, K., and O. Madsen. 1980. A comparative study of heuristics for a two-level routing-location problem. *European Journal of Operational Research,* 5, 378–387.

Jakobs, S. 1996. On genetic algorithms for the packing of polygons. *European Journal of Operational Research,* 88, 165–181.

Joint Publication 4-01.4. 2001. Department of the Chairman of the Joint Chiefs of Staff. *Joint tactics techniques and procedures for joint theater distribution*. Washington, DC.

Kelly, J.P., and J. Xu. 1998. *Technical Report*. Graduate School of Business, University of Colorado at Boulder.

Lambert, G.R. 2004. *A Tabu Search Approach to the Strategic Airlift Problem*. Ph.D. Dissertation, The University of Texas at Austin.

Lambert, G.R., J.W. Barnes, and D. Van Veldhuizen. 2007. A tabu search approach to the strategic airlift problem. *Military Operations Research,* 12(3), 59–72.

Laporte, G., 1988. *Location Routing Problems in Vehicle Routing Methods and Studies*, edited by B.C. Golden and A.A. Assad. Amsterdam: North-Holland, 163–198.

Li, Z., and V. Milenkovic. 1995. Compaction and separation algorithms for nonconvex polygons and their applications. *European Journal of Operational Research,* 84, 539–561.

Lim, A., B. Rodrigues, and Y. Wang. 2003. A multi-faced buildup algorithm for three-dimensional packing problems. *OMEGA*, 31 (6), 471–481.

Liu, N.C., and L.C. Chen. 1981. A new algorithm for container loading. *Compsac 81-5th International Computer Software and Applications Conference Papers,* IEEE, 292–299.

Liu, D., and H. Teng. 1999. An improved BL-algorithm for genetic algorithm of the orthogonal packing of rectangles. *European Journal of Operational Research,* 112, 413–420.

Lodi, A., S. Martello, and M. Monaci. 2002. Two-dimensional packing problems: A survey. *European Journal of Operational Research,* 141, 241–252.

Lodi, A., S. Martello, and D. Vigo. 1999a. Heuristic and metaheuristic approaches for a class of two-dimensional bin packing problems. *INFORMS Journal on Computing,* 11, 345–357.

Lodi, A., S. Martello, and D. Vigo. 1999b. Approximation algorithms for the oriented two-dimensional bin packing problem. *European Journal of Operational Research,* 112, 158–166.

Lodi, A., S. Martello, and D. Vigo. 2002a. Recent advance in two-dimensional bin packing problems. *Discrete Applied Mathematics,* 123, 379–396.

Lodi, A., S. Martello, and D. Vigo. 2002b. Heuristic algorithms for the three-dimensional bin packing problem. *European Journal of Operational Research,* 141, 410–420.

Loh, H.T., and A.Y.C. Nee. 1992. A packing algorithm for hexahedral boxes, *Proceedings of the Industrial Automation `92 Conference*, Singapore, 115–126.

Lourenco, H.R., J.P. Paixao, and R. Portugal. 1998. Metaheuristics for the bus-driver scheduling problem. *Economic Working Papers Series No. 304*, Universitat Pompeu Fabra, Spain.

Martello, S., D. Pisinger, and D. Vigo. 2000. The three-dimensional bin packing problem. *Operations Research*, 48 (2), 256–267.

Martello, S., and D. Vigo. 1998. Exact solution of the two-dimensional finite bin packing problem. *Management Science,* 44, 388–399.

McKinzie, K., and J.W. Barnes. 2004. A review of strategic mobility models supporting the defense transportation system. *Mathematical and Computer Modeling*, 39 (6–8), 839–868.

McKinzie, K., and J.W. Barnes. 2006. A tabu search approach to the strategic mobility mode selection problem (with K. McKinzie). *Air Force Journal of Logistics*, XXX, (3), 51–61.

McNabb, D.J. 2006. A message from General McNabb. *Airlift/Tanker Quarterly,* Fall, 10–13.

Min, H., V. Jayaram, and R. Srivastava. 1998. Combined location-routing problems: A synthesis and future research directions. *European Journal of Operational Research,* 10 (1), 10–23.

Mitrovic-Minic, S. 1998. Pickup and delivery problem with time windows: A survey No. SFU CMPT TR 1998-12.

Nagy, G., and S. Salhi. 1996. A nested location-routing heuristic method using route length estimation. *Studies in Locational Analysis*, 10(2), 109–127.

Nambier, J., L. Gelder, and L. Wassenhove. 1989. Plant location and vehicle routing in the Malaysian rubber smallholder sector: A case study. *European Journal of Operational Research,* 38 (38), 14–26.

Or, I., and W. Pierskalla. 1979. A transportation location – allocation model for the original blood banking. *AIIE Transactions,* 11, 86–95.

Perl, J., and M.S. Daskin. 1985. A warehouse location-routing problem. *Transportation Research,* 19B (5), 381–396.

Pisinger, D. 2002. Heuristics for the container loading problem. *European Journal of Operational Research*, 141 (2), 382–392.

Raidl, G.R., and G. Kodydek. 1998. Genetic algorithms for the multiple container packing problem. *Lecture Notes in Computer Science,* 1498, 875–884.

Reeves, C. 1995. *Modern Heuristic Techniques for Combinatorial Problems*. McGraw-Hill Book Company, London.

Roesener, A.G. 2006. *An Advanced Tabu Search Approach to the Airlift Loading Problem.* Ph.D. Dissertation, University of Texas at Austin.

Roesener, A.G., J.W. Barnes, J.T. Moore, and D. Van Veldhuizen. 2008. An advanced tabu search approach to the static airlift loading problem. *Military Operations Research*, submitted.

Romaine, J.M. 1999. *Solving the Multidimensional Multiple Knapsack Problem with Packing Constraints using Tabu Search*. Masters Thesis, Air Force Institute of Technology.

Sagan, B.E. 1991. *The Symmetric Group*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Rochat, Y., and E.D. Taillard. 1995. Probabilistic diversification and intensification in local search for vehicle routing. *Journal of Heuristics*, 1, 147–167.

Salhi, S., M. Fraser. 1996. An integrated heuristic approach for the combined location vehicle fleet mix problem. *Studies in Locational Analysis*, 8, 3–21.

Shen, Y., and R.S.K. Kwan. 2000. Tabu Search for Driver Scheduling. *Technical Report 2000.10*, University of Leeds.

Szykman, S., and J. Cagan. 1997. Constrained three-dimensional component layout using simulated annealing. *ASME Journal of Mechanical Design*, 119 (3), 28–35.

Terno, J., G. Scheithauer, U. Sommerweiß, and J. Riehme. 2000. An efficient approach for the multi-pallet loading problem. *European Journal of Operational Research*, 123 (2), 372–381.

Toth, P., and D. Vigo. 1997. Heuristic algorithms for the handicapped persons transportation problem. *Transportation Science,* 31 (1), 60–71.

TRADOC Pamphlet 525-3-90. 2002. US Army Operational and Organizational Plan for Maneuver Unit of Action, July.

Tuzun, D., and L. Burke. 1999. A two-phase tabu search approach to the location routing problem. *European Journal of Operational Research,* 116 (1), 87–99.

USTRANSCOM, 2005. United States Transportation Command. 9 November 2006. http://www.transcom.mil/organization.cfm.

Wiley, V. 2001. *The Aerial Fleet Refueling Problem*. Ph.D. Dissertation, The University of Texas at Austin.

Wu, Y.L., W. Huang, S. Lau, C.K. Wong, and G.H. Young. 2002. An effective quasi-human based heuristic for solving the rectangle packing problem. *European Journal of Operational Research*, 141, (2), 341–358.

Wu, T., C. Low, and J. Bai. 2002. Heuristic solutions to multi-depot location routing problems. *Computers and Operations Research,* 29 (10), 1393–1415.

7

New Advances in Solving the Weapon–Target Assignment Problem*

Ravindra K. Ahuja
University of Florida

Arvind Kumar
Innovative Scheduling, Inc.

Krishna C. Jha
Innovative Scheduling, Inc.

James B. Orlin
Massachusetts Institute of Technology

7.1 Introduction

The *weapon–target assignment (WTA) problem* arises in the modeling of combat operations where we wish to maximize the total expected damage caused to the enemy's target using limited number of weapons. In the WTA problem, there are n targets, numbered 1, 2, …, n, and m weapon types, numbered 1, 2, …, m. Let V_j denote the value of the target j, and W_i denote the number of weapons of type i available to be assigned to targets. Let p_{ij} denote the probability of destroying target j by a single weapon of type i. Hence $q_{ij}=1-p_{ij}$ denotes the probability of survival of target j if a single weapon of type i is assigned to it. Observe that if we assign x_{ij} number of weapons of type i to target j, then the survival probability of target j is given by $q_{ij}^{x_{ij}}$. A target may be assigned weapons of different types. The WTA problem is to determine the number of weapons x_{ij} of type i to be assigned to target j to minimize the total expected survival value of all targets. This problem can be formulated as the following nonlinear integer programming problem:

$$\text{Minimize} \sum_{j=1}^{n} V_j \left(\sum_{i=1}^{m} q_{ij}^{x_{ij}} \right) \tag{7.1a}$$

* The journal paper on which this chapter is based, won the "Koopman Prize" award as the best publication of the 2007 year by the Military Operations Research Society

subject to

$$\sum_{j=1}^{n} x_{ij} \leq W_i \text{, for all } i=1, 2, \ldots, m, \tag{7.1b}$$

$$x_{ij} \geq 0 \text{ and integer, for all } i=1, 2, \ldots, m, \text{ and for all } j=1, 2, \ldots, n. \tag{7.1c}$$

In the above formulation, we minimize the expected survival value of the targets while ensuring that the total number of weapons used is no more than those available. This formulation presents a simplified version of the WTA problem. In more practical versions, we may consider adding additional constraints, such as (i) lower and/or upper bounds on the number of weapons of type i assigned to a target j; (ii) lower and/or upper bounds on the total number of weapons assigned to target j; or (iii) a lower bound on the survival value of the target j. The algorithms proposed in this chapter can be easily modified to handle these additional constraints.

The weapon–target assignment problem has been widely studied by researchers over three decades. There are two types of the WTA problem: *static* and *dynamic*. In the static WTA problem, all the inputs to the problem are fixed; that is, all targets are known, all weapons are known, and all weapons engage targets in a single stage. The dynamic WTA problem is a generalization of the static WTA problem. This problem is a multistage problem where some weapons are engaged at the targets at a stage, the outcome of this engagement is assessed and strategy for the next stage is decided. The problem consists of a number of time stages and each time stage allows the defense to fire a subset of its weapons and observe perfectly the outcomes of all of the engagements of its weapons. With the feedback of this information, the defense can make better use of its weapons, since it will no longer engage targets which have already been destroyed. This is also called the *shoot-and-look* strategy. In real-life applications, solutions of the dynamic WTA problem are used in decision making. In this chapter, we study the static WTA problem; however, our algorithms can be used as important subroutines to solve the dynamic WTA problem. Henceforth, in this chapter, the static WTA problem is simply referred to as the WTA problem.

Research on the WTA problem dates back to the 1950s and 1960s where the modeling issues for the WTA problem were investigated (Manne 1958; Braford 1961; Day 1966). Lloyd and Witsenhausen (1986) established the NP-completeness of the WTA problem. Exact algorithms have been proposed to solve the WTA problem for the following special cases: (i) when all the weapons are identical (DenBroeder et al. 1958; Katter 1986); or (ii) when targets can receive at most one weapon (Chang et al. 1987; Orlin 1987). Castanon et al. (1987) have formulated the WTA problem as a nonlinear network flow problem; they relax the integrality constraints and the fractional assignments are fixed during a post-processing phase. Wacholder (1989) developed methods for solving the WTA problem using neural networks. The methods were tested on small scale problems only. Green et al. (1997) applied a goal programming-based approach to the WTA problem to meet other realistic preferences in a combat scenario. Metler and Preston (1990) have studied a suite of algorithms for solving the WTA problem efficiently, which is critical for real-time applications of the WTA problem. Maltin (1970), Eckler and Burr (1972), and Murphey (1999) provide comprehensive reviews of the literature on the WTA problem. Research to date on the WTA problem either solves the WTA problem for special cases or develops heuristics for the WTA problem. Moreover, since no exact algorithm is available to solve WTA problems, it is not known how accurate the solutions obtained by these heuristic algorithms are.

In this chapter, we propose exact and heuristic algorithms to solve the WTA problem using branch and bound techniques. Our branch and bound algorithms are the first implicit enumeration algorithms that can solve moderately sized instances of the WTA problem optimally. We also propose heuristic algorithms which generate near-optimal solutions within a few seconds of computational time. This chapter makes the following contributions:

- We formulate the WTA problem as an integer linear programming problem and as a generalized integer network flow problem on an appropriately defined network. The piecewise linear approximation of objective function of this formulation gives a lower bound on the optimal solution of the WTA problem. We describe this formulation in Section 7.2.1.

- We propose a minimum cost flow formulation that yields a different lower bound on the optimal solution of the WTA problem. This lower bound is, in general, not as tight as the bound obtained by the generalized integer network flow formulation described above, but it can be obtained in much less computational time. We describe this formulation in Section 7.2.2.
- We propose a third lower bounding scheme in Section 7.2.3 which is based on simple combinatorial arguments and uses a greedy approach to obtain a lower bound.
- We develop branch and bound algorithms to solve the WTA problem employing each of the three bounds described above. These algorithms are described in Section 7.3.
- We propose a very large-scale neighborhood (VLSN) search algorithm to solve the WTA problem. The VLSN search algorithm is based on formulating the WTA problem as a partition problem. The VLSN search starts with a feasible solution of the WTA problem and performs a sequence of *cyclic and path exchanges* to improve the solution. In Section 7.4, we describe a heuristic method that obtains an excellent feasible solution of the WTA problem by solving a sequence of minimum cost flow problems, and then uses a VLSN search algorithm to iteratively improve this solution.
- We perform extensive computational investigations of our algorithms and report these results in Section 7.5. Our algorithms solve moderately large instances (up to 80 weapons and 80 targets) of the WTA problem optimally and obtain almost optimal solutions of fairly large instances (up to 200 weapons and 200 targets) within a few seconds.

Though the focus of this chapter is to make contributions to a well known and unsolved operations research problem arising in defense, the chapter also makes some contributions that go beyond defense applications. We list some of these contributions next.

- Network cost flow problems are widely encountered in practice but applications of network flow problems to solve nonlinear integer programming problems are rather infrequent. This chapter illustrates an intuitive application of networks to this field. It shows that the generalized minimum cost flow and the (pure) minimum cost flow problems can be used to obtain lower bounds for nonlinear integer programming problems. It also describes a minimum cost flow based algorithm to obtain a near-optimal solution. These novel applications are likely to motivate similar applications for other nonlinear programming problems.
- Most previous applications of VLSN search techniques reported in the literature are for solving integer linear programming problems. This chapter shows that VLSN search algorithms can be applied to solve highly nonlinear integer programming problems as well. This work should lead to VLSN search algorithms for other integer nonlinear programming problems as well.

7.2 Lower-bounding Schemes

In this section, we describe three lower bounding schemes for the WTA problem, using linear programming, integer programming, minimum cost flow problem, and a combinatorial method. These three approaches produce lower bounds with different values and have different running times.

7.2.1 A Lower-bounding Scheme using an Integer Generalized Network Flow Formulation

In this section, we formulate the WTA problem as an integer-programming problem with a convex objective function value. This formulation is based on a result reported by Manne (1958) who attributed it to Dantzig (personal communications).

In Formulation 7.1, let $s_j = \prod_{i=1}^{m} q_{ij}^{x_{ij}}$. Taking logarithms on both sides, we obtain, $\log(s_j) = \sum_{i=1}^{m} x_{ij}\log(q_{ij})$ or $-\log(s_j) = \sum_{i=1}^{m} x_{ij}\left(-\log(q_{ij})\right)$. Let $y_j = -\log(s_j)$ and $d_{ij} = -\log(q_{ij})$. Observe that since $0 \le q_{ij} \le 1$, we have $d_{ij} \ge 0$. Then $y_j = \sum_{i=1}^{m} d_{ij}x_{ij}$. Also observe that $\prod_{i=1}^{m} q_{ij}^{x_{ij}} = 2^{-y_j}$. By introducing the terms d_{ij} and y_j in Formulation 7.1, we get the following formulation:

$$\text{Minimize} \sum_{j=1}^{n} V_j 2^{-y_j} \tag{7.2a}$$

subject to

$$\sum_{j=1}^{n} x_{ij} \le W_i \text{ for all } i = 1, 2, \ldots, m, \tag{7.2b}$$

$$\sum_{i=1}^{m} d_{ij}x_{ij} = y_j \text{ for all } j = 1, 2, \ldots, n, \tag{7.2c}$$

$$x_{ij} \ge 0 \text{ and integer for all } i = 1, \ldots, m \text{ and for all } j = 1, \ldots, n, \tag{7.2d}$$

$$y_j \ge 0 \text{ for all } j = 1, 2, \ldots, n. \tag{7.2e}$$

Observe that Formulation 7.2 is an integer programming problem with separable convex objective functions. This integer program can also be viewed as an integer generalized network flow problem with convex flow costs. Generalized network flow problems are flow problems where flow entering an arc may be different than the flow leaving the arc (see, for example, Ahuja et al. 1993). In a generalized network flow problem, each arc (i, j) has an associated multiplier γ_{ij} and the flow x_{ij} becomes $\gamma_{ij}x_{ij}$ as it travels from node i to node j. The Formulation 7.2 is a generalized network flow problem on the network shown in the Figure 7.1. Next we give some explanations of this formulation.

The network contains m weapon nodes, one node corresponding to each weapon type. The supply at node i is equal to the number of weapons available, W_i, for the weapon type i. The network contains n target nodes, one node corresponding to each target, and there is one sink node t. The supplies/demands of target nodes are zero. We now describe the arcs in the network. The network contains an arc connecting each weapon node to each target node. The flows on these arcs are given by x_{ij}, representing the number of weapons of type i assigned to the target j. The multipliers for these arcs are d_{ij}'s. Since there is no cost coefficient for x_{ij}'s in the objective function, the cost of flow on these arcs is zero. The network contains an arc from each of the target nodes to the sink node t. The flow on arc (j, t) is given by y_j and

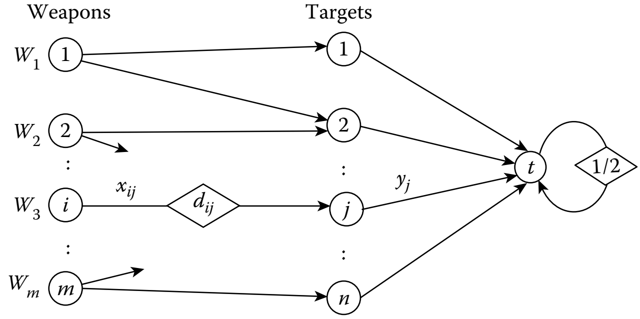


FIGURE 7.1 Formulating the WTA problem as an integer generalized network flow problem.

the cost of flow on this arc is $V_j 2^{-y_j}$. Finally, there is a loop arc (t, t) incident on node t with multiplier ½. The flow on this arc is equal to the double of the sum of the flow on incoming arcs from the target nodes. This constraint is required to satisfy the mass balance constraints at node t.

In Formulation 7.2, the cost of the flow in the network equals the objective Function 7.2a, the mass balance constraints of weapon nodes are equivalent to the Constraint 7.2b, and mass balance constraints of target nodes are equivalent to the Constraint 7.2c. It follows that an optimal solution of the above generalized network flow problem will be an optimal solution of the WTA problem.

The generalized network flow Formulation 7.2 is substantially more difficult than the standard generalized network flow problem (see, Ahuja et al. 1993) since the flow values x_{ij}'s are required to be integer numbers (instead of real numbers) and the costs of flows on some arcs is a convex function (instead of a linear function). We will approximate each convex function by a piecewise linear convex function so that the optimal solution of the modified formulation gives a lower bound on the optimal solution of the generalized Formulation 7.2.

We consider the cost function $V_j 2^{-y_j}$ at values y_j that are integer multiples of a parameter $p > 0$, and draw tangents of $V_j 2^{-y_j}$ at these values. Let $F_j(p, y_j)$ denote the upper envelope of these tangents. It is easy to see that the function $F_j(p, y_j)$ approximates $V_j 2^{-y_j}$ from below and for every value of y_j provides a lower bound on $V_j 2^{-y_j}$. Figure 7.2 shows an illustration of this approximation.

Thus, in Formulation 7.2 if we replace the objective function (Equation 7.2a) by the following objective function,

$$\sum_{j=1}^{n} F_j(p, y_j), \tag{7.2a$'$}$$

we obtain a lower bound on the optimal objective function of Equation 7.2a. Using this modified formulation, we can derive lower bounds in two ways:

LP-based lower bounding scheme: Observe that the preceding formulation is still an integer programming problem because arc flows x_{ij}'s are required to be integer valued. By relaxing the integrality of the x_{ij}'s, we obtain a mathematical programming problem with linear constraints and piecewise linear convex objective functions. It is well-known (see, Murty 1976) that linear programs with piecewise linear convex functions can be transformed to linear programs by introducing a variable for every linear

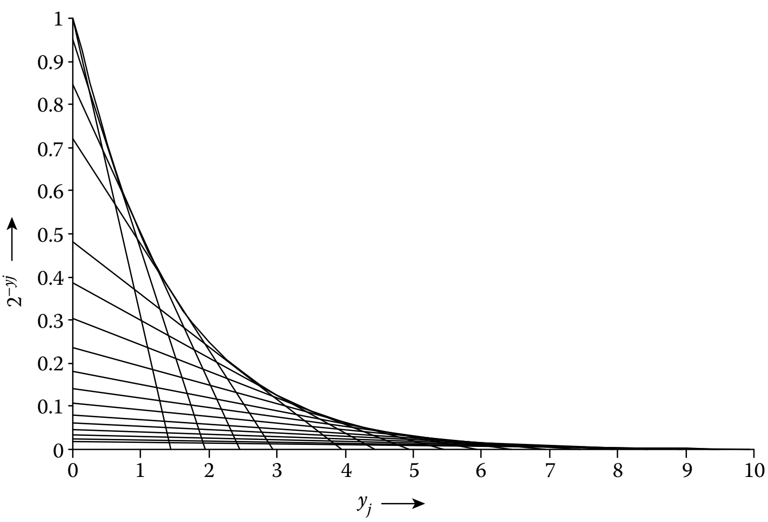


FIGURE 7.2 Approximating a convex function by a lower envelope of linear segments.

segment. We can solve this linear programming problem to obtain a lower bound for the WTA problem. Our computational results indicate the lower bounds generated by this scheme are not very tight, and therefore, we do not include this scheme in our discussions.

MIP-based lower bounding scheme: In this scheme, we do not relax the integrality of the x_{ij}'s, which keeps the formulation as an integer programming formulation. However, we transform the piecewise linear convex functions to linear cost functions by introducing a variable for every linear segment. We then use cutting plane methods to obtain a lower bound on the optimal objective function value. We have used the built-in routines in the software CPLEX 8.0 to generate Gomory cuts and mixed integer rounding cuts to generate fairly tight lower bounds for the WTA problem. It can be observed that for a sufficiently high number of segments and when given enough time to converge, this method will converge to an (near) optimal solution of the original problem.

We summarize the discussion in this section as follows:

Theorem 7.1. *The generalized network flow formulation gives a lower bound on the optimal objective function value for the WTA problem.*

7.2.2 A Minimum Cost Flow-based Lower Bounding Scheme

The objective function of the WTA problem can also be interpreted as maximizing the expected damage to the targets. In this section, we develop an upper bound on the expected damage to the targets. Subtracting this upper bound on the expected damage from the total value of the targets ($\sum_{j=1}^{n} V_j$) will give us a lower bound on the minimum survival value. We formulate the problem of maximizing the damage to targets as a maximum cost flow problem. The underlying network G for the maximum cost flow formulation is shown in Figure 7.3.

This network has three layers of nodes. The first layer contains a supply node i for every weapon type i with supply equal to W_i. We denote these supply nodes by the set N_1. The second layer of nodes, denoted by the set N_2, contains nodes corresponding to targets, but each target j is represented by several nodes

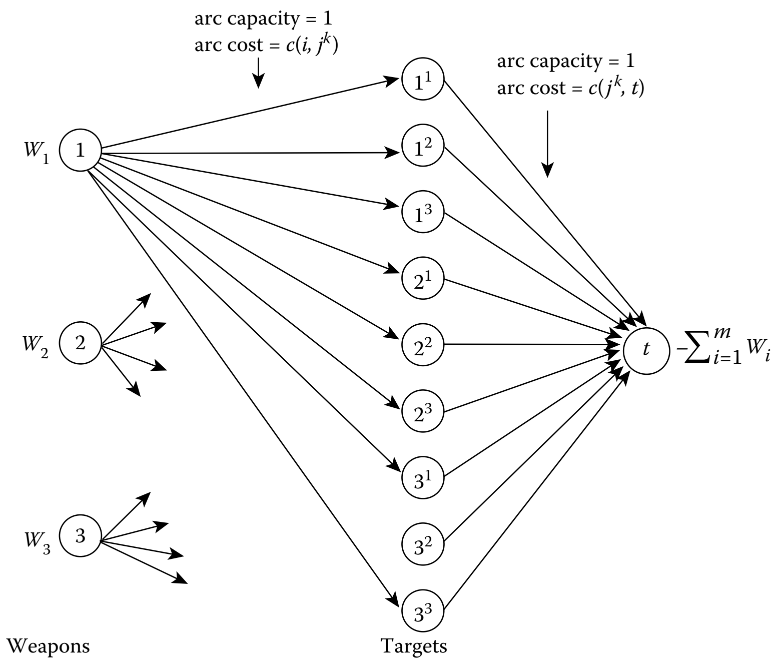


FIGURE 7.3 A network flow formulation of the WTA problem.

$j^1, j^2, \ldots, j^k$, where k is the maximum number of weapons that can be assigned to target j. A node j^p represents the p^{th} weapon striking the target j, i.e. $p-1$ number of weapons has already been assigned to the target j. For example, the node labeled 3^1 represents the event of the first weapon being assigned to target 3, the node labeled 3^2 represents the event of the second weapon being assigned to target 3, and so on. All nodes in the second layer have zero supplies/demands. Finally, the third layer contains a singleton node t with demand equal to $\sum_{i=1}^{m} W_i$.

We now describe the arcs in this network. The network contains an arc (i, j^k) for each node $i \in N_1$ and each node $j^k \in N_2$; this arc represents the assignment of a weapon of type i to target j as the k^{th} weapon. This arc has a unit capacity. The network also contains an arc (j^k, t) with unit capacity for each node $j^k \in N_2$.

We call a flow x in this network a *contiguous flow* if it satisfies the property that if $x(i, j^k) = 1$, then $x(i, j^l) = 1$ for all $l = 1, 2, \ldots, k-1$. In other words, the contiguous flow implies that weapon i is assigned to target j as the k^{th} weapon provided that $(k-1)$ weapons have already been assigned to it. The following observation directly follows from the manner we have constructed the network G:

Observation 7.1 *There is one-to-one correspondence between feasible solutions of the WTA problem and contiguous flows in G.*

While there is a one-to-one correspondence between feasible solutions, it is not a cost-preserving correspondence if we require costs to be linear. We instead provide linear costs that will overestimate the true nonlinear costs. We define our approximate costs next.

The arc (i, j^k) represents the assignment of a weapon of type i to target j as the k^{th} weapon. If $k = 1$, then the cost of this arc is the expected damage caused to the target,

$$c(i, j^1) = V_j(1 - q_{ij}) \tag{7.3}$$

which is the difference between the survival value of the target before strike (V_j) and the survival value of the target after strike ($V_j q_{ij}$). Next, consider the cost $c(i, j^2)$ of the arc (i, j^2) which denotes the change in the survival value of target j when weapon i is assigned to it as the second weapon. To determine this, we need to know the survival value of target j before weapon i is assigned to it. But this cost depends upon which weapon was assigned to it as the first weapon. The first weapon striking target j can be of any weapon type $1, 2, \ldots, m$ and we do not know its type a priori. Therefore, we cannot determine the cost of the arc (i, j^2). However, we can determine an upper bound on the cost of the arc (i, j^2). Next we will derive the expression for the cost of the arc (i, j^k) which as a special case includes (i, j^2).

Suppose that the first $(k-1)$ weapons assigned to target j are of weapon types $i_1, i_2, \ldots, i_{k-1}$, and suppose that the type of the k^{th} assigned weapon is of type i. Then, the survival value of target j after the first $(k-1)$ weapons is $V_j q_{i_1 j} q_{i_2 j} \ldots q_{i_{k-1} j}$ and the survival value of the target j after k weapons is $V_j q_{i_1 j} q_{i_2 j} \ldots q_{i_{k-1} j} q_{ij}$. Hence, the cost of the arc (i, j^k) is the difference between the two terms, which is

$$c(i, j^k) = V_j q_{i_1 j} q_{i_2 j} \ldots q_{i_{k-1} j}(1 - q_{ij}). \tag{7.4}$$

Let $q_j^{\max} = \max\{q_{ij}: 1, 2, \ldots, m\}$.Then, we can obtain an upper bound on $c(i, j^k)$ by replacing each q_{ij} by $q_j^{\max}$. Hence, if we set

$$c(i, j^k) = V_j (q_j^{\max})^{k-1}(1 - q_{ij}), \tag{7.5}$$

we get an upper bound on the total destruction on assigning weapons to targets. It directly follows from Equation 7.5 that

$$c(i, j^1) > c(i, j^2) > \ldots > c(i, j^{k-1}) > c(i, j^k), \tag{7.6}$$

which implies that the optimal maximum cost flow in the network G will be a contiguous flow. It should be noted here that since this is a maximization problem, we solve it by first multiplying all arc costs by -1 and then using any minimum cost flow algorithm. Let z^* represent the upper bound on the destruction caused to targets after all the assignments obtained by solving this maximum cost flow problem. Then, the lower bound on the objective function of Formulation 7.1 is $\sum_{j=1}^{n} V_j - z^*$.

We can summarize the preceding discussion as follows:

Theorem 7.2. *If z^* is the optimal objective function value of the maximum cost flow problem in the network G, then $\sum_{j=1}^{n} V_j - z^*$ is a lower bound for the weapon–target assignment problem.*

7.2.3 Maximum Marginal Return-based Lower Bounding Scheme

In this section, we describe a different relaxation that provides a valid lower bound for the WTA problem. This approach is based on the underestimation of the survival of a target when hit by a weapon as we assume that every target is hit by the best weapons.

Let $q_j^{\min}$ be the survival probability for target j when hit by the weapon with the smallest survival probability, i.e. $q_j^{\min} = \min\{q_{ij}: i = 1, 2, \ldots, m\}$. Replacing the term q_{ij} in Formulation 7.1 by $q_j^{\min}$, we can formulate the WTA problem as follows:

$$\text{Minimize} \sum_{j=1}^{n} V_j \prod_{i=1}^{m} (q_j^{\min})^{x_{ij}} \tag{7.7a}$$

subject to

$$\sum_{j=1}^{n} x_{ij} \le W_i, \text{ for all } i = 1, 2, \ldots, m, \tag{7.7b}$$

$$x_{ij} \ge 0 \text{ and integer for all } i = 1, 2, \ldots, m \text{ and for all } j = 1, 2, \ldots, n. \tag{7.7c}$$

Let $x_j = \sum_{i=1}^{m} x_{ij}$, and if we let $g_j(x_j) = V_j (q_j^{\min})^{x_j}$, then we can rewrite Equation 7.7 as:

$$\text{Minimize} \sum_{j=1}^{n} g_j(x_j) \tag{7.8a}$$

subject to

$$\sum_{j=1}^{n} x_{ij} \le W_i, \text{ for all } i = 1, 2, \ldots, m, \tag{7.8b}$$

$$\sum_{i=1}^{n} x_{ij} = x_j \text{ for all } i = 1, 2, \ldots, m, \tag{7.8c}$$

$$x_{ij} \ge 0 \text{ and integer for all } i = 1, 2, \ldots, m \text{ and for all } j = 1, 2, \ldots, n. \tag{7.8d}$$

It is also possible to eliminate the variables x_{ij} entirely. If we let $W = \sum_{i=1}^{m} W_i$, Formulation 7.8 can be rewritten as an equivalent integer program (Equation 7.9):

$$\text{Minimize} \sum_{j=1}^{n} g_j(x_j) \tag{7.9a}$$

subject to

$$\sum_{j=1}^{n} x_j \leq W, \tag{7.9b}$$

$$x_j \geq 0 \text{ and integer for all } j = 1, 2, \ldots, n. \tag{7.9c}$$

It is straightforward to transform a solution for Equation 7.9 into one for Equation 7.8 since all weapon types are identical in Formulation 7.8.

Observe that the Formulations 7.1 and 7.7 have the same constraints; hence, they have the same set of solutions. However, in the Formulation 7.7, we have replaced each q_{ij} by $q_j^{\min} = \min\{q_{ij}: i = 1, 2, \ldots, m\}$. Noting that the optimal solution value for Equations 7.7 through 7.9 are all identical, we get the following result:

Theorem 7.3 *An optimal solution value for Equation 7.9 is a lower bound for the WTA problem.*

Integer program (Equation 7.9) is a special case of the knapsack problem in which the separable costs are monotone decreasing and concave. As such, it can be solved using the greedy algorithm. In the following, assigned(j) is the number of weapons assigned to target j, and value(i, j) is the incremental cost of assigning the next weapon to target j (Figure 7.4).

This lower bounding scheme is in fact a variant of a popular algorithm to solve the WTA problem, which is known as the *maximum marginal return algorithm*. In this algorithm, we always assign a weapon with maximum improvement in the objective function value. This algorithm is a heuristic algorithm to solve the WTA problem, but is known to give an optimal solution if all weapons are identical.

We now analyze the running time of our lower bounding algorithm. In this algorithm, once the value (*value(j)*) of each target is initialized, we assign W weapons iteratively and in each iteration we: (i) identify target with minimum *value(j)*, and (ii) decrease the value of identified target. If we maintain a Fibonacci heap of targets with their values as their keys, then the running time of our algorithm in each iteration is O(1) for identifying the minimum value target plus O(1) in updating the heap when the value of identified target is decreased (Fredman and Tarjan 1984). As the algorithm performs W iterations, the running time of the Fibonacci heap implementation of our maximum marginal return algorithm is O(W). In our implementation, we used binary heap in place of Fibonacci heap, which runs in O(W log n) time and is still fast enough for the problem sizes that we considered.

```
algorithm combinatorial-lower-bounding;
begin
   for j := 1 to n do
   begin
      assigned(j):= 0;
      value(j):= g_j(assigned(j) + 1) - g_j(assigned(j));
   end
   for i = 1 to m do
   begin
      find j corresponding to the minimum value(j);
      assigned(j):= assigned(j) + 1;
      value(j):= g_j(assigned(j)+1) - g_j(assigned(j));
   end;
end.
```

FIGURE 7.4 Combinatorial lower bounding algorithm.

7.3 A Branch and Bound Algorithm

We developed and implemented a branch and bound algorithm and experimented it for three lower bounding strategies. A branch and bound algorithm is characterized by the branching, lower bounding, and search strategies. We now describe these strategies for our approaches.

Branching strategy: For each node of the branch and bound tree, we find the weapon–target combination which gives the best improvement and set the corresponding variable as the one to be branched on next. Ties are broken arbitrarily. To keep the memory requirement low, the only information we store at any node is which variable we branch on at that node and the lower and upper bounds at the node. The upper bound at a node is the sum of remaining survival value of all the targets. To recover the partial solution associated with a node of the branch and bound tree, we trace back to the root of the tree.

Lower bounding strategy: We used the three lower bounding strategies: (i) generalized network flow (Section 7.2.1); (ii) minimum cost flow (Section 7.2.2); and (iii) maximum marginal return (Section 7.2.3). We provide a comparative analysis of these bounding schemes in Section 7.5.

Search strategy: We implemented both the breadth-first and depth-first search strategies. We found that for smaller size problems (i.e. up to ten weapons and ten targets), breadth-first strategy gave overall better results, but for larger problems, depth-first search had a superior performance. We report the results for the depth-first search in Section 7.5.

7.4 A Very Large-scale Neighborhood Search Algorithm

In the previous section, we described a branch and bound algorithm for the WTA problem. This algorithm is the first exact algorithm that can solve moderate size instances of the WTA problem in reasonable time. Nevertheless, there is still a need for heuristic algorithms which can solve large-scale instances of the WTA problems. In this section, we describe a neighborhood search algorithm for the WTA problem which has exhibited excellent computational results. A neighborhood search algorithm starts with a feasible solution of the optimization problem and successively improves it by replacing the solution by an improved neighbor until converges to a locally optimal solution. The quality of the locally optimal solution depends both upon the quality of the starting feasible solution and the structure of the neighborhood; that is, how we define the neighborhood of a given solution. We next describe the method we used to construct the starting feasible solution followed by our neighborhood structure. This algorithm is an application of a very large-scale neighborhood (VLSN) search to the WTA problem. A VLSN search algorithm is a neighborhood search algorithm where the size of the neighborhood is very large and we use some implicit enumeration algorithm to identify an improved neighbor.

7.4.1 A Minimum Cost Flow Formulation-based Construction Heuristic

We developed a construction heuristic which solves a sequence of minimum cost flow problems to obtain an excellent solution of the WTA problem. This heuristic uses the minimum cost flow formulation shown in Figure 7.3, which we used to determine a lower bound on the optimal solution of the WTA problem. Recall that in this formulation, we define the arc costs (i, j^1), (i, j^2), …, (i, j^k), which, respectively, denote the cost of assigning the first, second and k^{th} weapon of type i to target j. Also recall that only the cost of the arc (i, j^1) was computed correctly, and for the other arcs, we used a lower bound on the cost. We call the arcs whose costs are computed correctly as *exact-cost arcs*, and the rest of the arcs as *approximate-cost arcs*.

This heuristic works as follows. We first solve the minimum cost flow problem with respect to the arc costs as defined earlier. In the optimal solution of this problem, exact-cost arcs as well as approximate-cost arcs may carry positive flow. We next fix the part of the weapon–target assignment corresponding to the flow on the exact-cost arcs and remove those arcs from the network. In other words, we construct a partial solution for the weapon–target assignment by assigning weapons only for exact-cost arcs. After fixing this partial assignment, we again compute the cost of each arc. Some previous approximate-cost arcs will now become exact-cost arcs. For example, if we set the flow on arc (i, j^1) equal to 1, we know that that weapon i is the first weapon striking target j, and hence we need to update the costs of the arcs (l, j^k) for all $l = 1, 2, \ldots, m$ and for all $k \geq 2$. Also observe that the arcs (l, j^2) for all $l = 1, 2, \ldots, m$ now become exact cost arcs. We next solve another minimum cost flow problem and again fix the flow on the exact-cost arcs. We recompute arc costs, make some additional arcs exact-cost, and solve another minimum flow problem. We repeat this process until all weapons are assigned to the targets.

We tried another modification in the minimum cost flow formulation which gave better computational results. The formulation we described determines the costs of approximate-cost arcs assuming that the worst weapons (with the largest survival probabilities) are assigned to targets. However, we observed that in any near-optimal solution, the best weapons are assigned to the targets. Keeping this observation in mind, we determine the costs of valid arcs assuming that the best weapons (with the smallest survival probabilities) are assigned to targets. Hence, the cost of the arc (i, j^k), which is $c(i, j^k) = V_j q_{i_1 j} q_{i_2 j} \ldots q_{i_{k-1} j}(1 - q_{ij})$ is approximated by $c(i, j^k) = V_j\, [q_{min}(j)]^{k-1}(1 - q_{ij})$. Our experimental investigation shows that this formulation generates better solutions compared to the previous formulation. We present computational results of this formulation in Section 7.5.

7.4.2 The VLSN Neighborhood Structure

The VLSN search heuristic uses *multiexchange neighborhood structures* that can be regarded as generalizations of the two-exchange neighborhood structures. This neighborhood structure is based on the multiexchange exchange neighborhood structure developed by Thompson and Psaraftis (1993) and Thompson and Orlin (1989). Glover (1996) refers to this type of exchange as *ejection chains* (see also, Rego and Roucairol 1996; Kelley and Xu 1996). Thompson and Psaraftis (1993) and Gendreau et al. (1998) have used this neighborhood structure to solve vehicle routing problems and obtained impressive results. We refer the reader to the paper by Ahuja et al. (2002) for an overview of VLSN search algorithms. In this chapter, we use multiexchange neighborhood structures to solve the WTA problem.

The WTA problem can be conceived of as a partition problem defined as follows. Let $\boldsymbol{S} = \{a_1, a_2, a_3, \ldots, a_n\}$ be a set of n elements. The partition problem is to partition the set $\boldsymbol{S}$ into the subsets $S_1, S_2, S_3, \ldots, S_K$ such that the cost of the partition is minimal, where the cost of the partition is the sum of the cost of each part. The WTA problem is a special case of the partition problem where the set of all weapons is partitioned into n subsets $S_1, S_2, \ldots, S_n$, and subset j is assigned to target j, $1 \leq j \leq n$. Thompson and Orlin (1989) and Thompson and Psaraftis (1993) proposed a VLSN search approach for partitioning problems which proceeds by performing *cyclic exchanges*. Ahuja et al. (2001, 2003) proposed further refinements of this approach and applied it to the capacitated minimum spanning tree problem. We will present a brief overview of this approach when applied to the WTA problem.

Let $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$ denote a feasible solution of the WTA problem where the subset S_j, $1 \leq j \leq n$, denotes the set of weapons assigned to target j. Our neighborhood search algorithm defines neighbors of the solution $\boldsymbol{S}$ as those solutions that can be obtained from $\boldsymbol{S}$ by performing *multiexchanges*. A *cyclic multiexchange* is defined by a sequence of weapons i_1-i_2-i_3-…-i_r-i_1 where the weapons $i_1, i_2, i_3, \ldots, i_r$ belong to different subsets S_j's. Let $t(i_1), t(i_2), t(i_3), \ldots, t(i_r)$, respectively, denote the targets to which weapons $i_1, i_2, i_3, \ldots, i_r$, are assigned. The *cyclic multiexchange* i_1- i_2- i_3-…-i_r- i_1 represents that weapon i_1 is reassigned from target $t(i_1)$ to target $t(i_2)$, weapon i_2 is reassigned from target $t(i_2)$ to target $t(i_3)$, and so on, and finally weapon t_r is reassigned from target $t(i_r)$ to target $t(i_1)$. In Figure 7.5a, we show a cyclic multiexchange involving four weapons (i_1, i_2, i_3, and i_4). We can similarly define a *path multiexchange* by a sequence of weapons i_1- i_2- i_3-…- i_r which differs from the *cyclic multiexchange* in the sense that the last weapon i_r is not reassigned and remains assigned to target $t(i_r)$. For example, in Figure 7.5b, weapon i_1

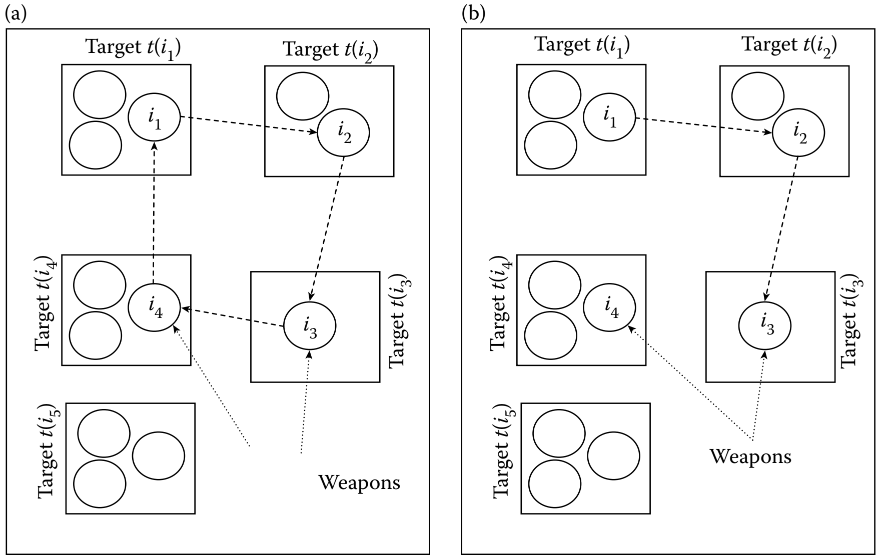


FIGURE 7.5 (a) An example of a cyclic multi-exchange; (b) an example of a path multi-exchange.

leaves target $t(i_1)$ and gets assigned to target $t(i_2)$, weapon i_2 leaves target $t(i_2)$ and gets assigned to target $t(i_3)$, and no weapon leaves target $t(i_3)$ as a result of path multiexchange.

The number of neighbors in the multiexchange neighborhood is too large to be enumerated explicitly. However, using the concept of *improvement graph*, a profitable multiexchange can be identified using network algorithms. The improvement graph $G(\boldsymbol{S})$ for a given feasible solution $\boldsymbol{S}$ of the WTA problem contains a node r corresponding to each weapon r and contains an arc (r, l) between every pair of nodes r and l with $t(r) \neq t(l)$. The arc (r, l) signifies the fact that weapon r is reassigned to target (say j) to which weapon l is currently assigned and weapon l is unassigned from its current target j. The cost of this arc (say c_{rl}), is set equal to the change in the survival value of the target j due to reassignments. Let V_j' denote the survival value of the target j in the current solution. Then, the cost of the arc (r, l) is $c_{rl} = V_j'\left((q_{rj}/q_{lj})-1\right)$. The terms V_j'/q_{lj} and $V_j'\ q_{rj}$ in previous expression, respectively, denote that weapon l leaves the target j and weapon r gets assigned to it. We say that a directed cycle $W = i_1$- i_2- i_3-…- i_k- i_1 in $G(\boldsymbol{S})$ is *subset-disjoint* if each of the weapons $i_1, i_2, i_3, \ldots, i_k$ is assigned to a different target. Thompson and Orlin (1989) showed the following result:

Lemma 7.1. *There is a one-to-one correspondence between multiexchanges with respect to **S** and directed subset-disjoint cycles in G(**S**) and both have the same cost.*

This lemma allows us to solve the WTA problem using the neighborhood search algorithm illustrated in Figure 7.6.

We now give some details of the VLSN search algorithm. We obtain the starting feasible solution $\boldsymbol{S}$ by using the minimum cost flow based heuristic described in Section 7.4.1. The improvement graph $G(\boldsymbol{S})$ contains W nodes and $O(W^2)$ arcs and the cost of all arcs can be computed in $O(W^2)$ time. We use a dynamic programming based algorithm (Ahuja et al. 2003) to obtain subset-disjoint cycles. This algorithm first looks for profitable two-exchanges involving two targets only; if no profitable two-exchange is found, it looks for profitable three-exchanges involving three targets, and so on. In each iteration of a neighborhood search, the algorithm either finds a profitable multiexchange or terminates when it is unable to find a multiexchange involving k targets (we set $k=5$). In the former case, we improve the current solution, and in the latter case we declare the current solution to be locally optimal and stop. The running time of the dynamic programming algorithm is $O(W^2 2^k)$ per iteration, and is typically much faster since most cyclic exchanges found by the algorithm are two-exchanges, i.e., $k=2$.

```
algorithm WTA-VLSN search;
begin
     obtain a feasible solution S of the WTA problem;
     construct the improvement graph G(S);
     while G(S) contains a negative cost subset-disjoint cycle do
     begin
          obtain a negative cost subset-disjoint cycle W in G(S);
          perform the multi-exchange corresponding to W;
          update S and G(S);
     end;
end;
```

FIGURE 7.6 The VLSN search algorithm for the WTA problem.

7.5 Computational Results

We implemented each of the algorithms described in the previous section and extensively tested them. We tested our algorithms on randomly generated instances as data for the real-life instances was classified. We generated the data in the following manner. We generated the target survival values V_j's as uniformly distributed random numbers in the range 25–100. We generated the kill probabilities for weapons–target combinations as uniformly distributed random numbers in the range 0.60–0.90. We performed all our tests on a 2.8-GHz Pentium 4 Processor computer with 1 GB RAM PC. In this section, we present the results of these investigations.

7.5.1 Comparison of the Lower Bounding Schemes

In our first investigation, we compared the tightness of the lower bounds generated by the lower bounding algorithms developed by us. We have proposed three lower bounding schemes described in Section 7.2: (i) the generalized network flow based lower bounding scheme; (ii) the minimum cost based lower bounding scheme; and (iii) the maximum marginal return based lower bounding scheme. The computational results of these three lower bounding schemes are given in Table 7.1. For each of these schemes, the first column gives the % gap from the optimal objective function value and the second column gives the time taken to obtain the bound. For the generalized network flow scheme, we used a piecewise approximation of the objective function and then solved the problem using CPLEX 8.0 MIP solver. For the piecewise approximation of the objective function, we did some sensitivity analysis for the number of segments used. We found that with lesser number of segments, the lower bounds achieved are not very good, and the algorithm enumerates too many candidate problems resulting in high running times. However, if we use more segments, then lower bounds are better but the running time at each node increases, thereby again increasing the total running time of the algorithm. Thus, one needs to balance these two conflicting considerations to find the overall best number of segments. We experimented with 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, and 200 segments. By empirical testing, we found that 50 segments gives the best overall performance. Thus, for all our experimentation, we used 50 segments.

We can derive the following observations from Table 7.1:

(i) The generalized network flow based lower bounding scheme gives the tightest lower bounds but also takes the maximum computational time.
(ii) The minimum cost flow based lower bounding scheme gives fairly tight lower bounds when the number of weapons (W) is less than or equal to the number of targets (n). It can be recalled that in the minimum cost flow formulation, the cost of arcs assigning first weapon to a target are

TABLE 7.1 Comparison of the Three Lower Bounding Schemes

# of Weapons	# of Targets	Generalized Network Flow Scheme		Minimum Cost Flow Scheme		Maximum Marginal Return Scheme	
		% Gap	Time (sec)	% Gap	Time (sec)	% Gap	Time (sec)
5	5	0.21	0.016	1.66	<0.001	10.61	<0.001
10	10	0.12	0.031	0.00	<0.001	11.01	<0.001
10	20	0.04	0.062	0.00	<0.001	1.45	<0.001
20	10	0.53	0.156	21.32	<0.001	19.00	<0.001
20	20	0.25	0.109	1.32	<0.001	6.40	<0.001
20	40	0.04	0.296	0.00	<0.001	1.57	<0.001
40	10	2.12	0.609	42.41	<0.001	46.89	<0.001
40	20	0.45	0.359	25.52	0.015	13.53	<0.001
40	40	0.11	0.703	1.63	0.015	3.05	<0.001
40	80	0.03	1.812	0.00	0.046	0.88	<0.001

exact. We also observed that when $W<n$, then the number of weapons assigned to most targets is at most one in the optimal solution. Therefore, this scheme gives very tight lower bounds when $W<n$.

(iii) The maximum marginal return based lower bounding scheme is very efficient and running times are quite low. However, the high optimality gap for maximum marginal return demonstrates that a simple greedy method is not very efficacious for the WTA problem.

7.5.2 Comparison of Three Lower Bounding Scheme in the Branch and Bound Algorithm

We developed a branch and bound algorithm using the three lower bounding schemes and one hybrid algorithm. The hybrid algorithm computes lower bounds using both the minimum cost flow based and the maximum marginal return based lower bounding schemes and uses the better of these two bounds. Table 7.2 gives the results of these algorithms. The cells containing "–" mean that our branch and bound algorithm could not found optimal solution even after running for more than 48 hours. We observe that the branch and bound algorithm using the MIP based lower bounding gives the most consistent results and is able to solve the largest size problems (containing 80 weapons and 80 targets). We also find that the hybrid algorithm also gives excellent results for those instances where the number of weapons is less than or equal to the number of targets.

7.5.3 Performance of the VLSN Search Algorithm

We now present computational results of the minimum cost flow based construction heuristic and the VLSN search algorithm. In Table 7.3, we give the objective function values of the solutions obtained by the construction heuristic and the improved values when VLSN search algorithm is applied to these solutions. For small instances, the optimal values were determined using the branch and bound method. For larger instances, the branch and bound algorithm could not be executed till optimality. However, observe that the minimum of the lower bounds of the active (node not pruned yet) nodes gives an overall lower bound on the objective function. We used this value to compute the optimality gap.

We believe that the success of VLSN search on the WTA problem can be attributed to three factors. First, the WTA problem is unconstrained in nature and VLSN approaches are very efficacious in solving unconstrained partition problems. Second, we search the solution space for up to five exchanges. This implies that we consider neighborhoods of size=(no-of-weapons)^5. For example, in the case of

TABLE 7.2 Comparison of Variations in Lower Bounding Schemes in Branch and Bound Algorithm

		Generalized Network Flow-based B&B Algorithm		Minimum Cost Flow Based B&B Algorithm		Maximum Marginal Return-based B&B Algorithm		Hybrid Algorithm	
# of Weapons	# of Targets	Nodes Visited	Time (sec)	Nodes Visited	Time (sec)	Nodes Visited	Time (sec)	Nodes Visited	Time (sec)
5	5	15	0.14	11	<0.001	23	<0.001	11	<0.001
10	10	29	0.56	1	<0.001	181	<0.001	1	<0.001
10	20	23	0.83	1	<0.001	83	<0.001	1	0.015
20	10	101	7.27	–	–	2,8611	1.34	20,251	2.52
20	20	109	6.56	2,383	4.39	15936	0.94	1,705	2.50
20	40	105	16.58	1	<0.001	111,603	10.14	1	0.015
40	10	1,285	327.27	–	–	–	–	–	–
40	20	205	35.19	–	–	~108	13,651.9	~107	25,868.9
40	40	211	50.96	~106	10,583.62	~106	943.03	38,3275	1,891.83
40	80	385	235.41	1	0.031	–	–	1	0.031
80	40	117,227	43,079.55	–	–	–	–	–	–
80	80	44905	58,477.31	–	–	–	–	–	–
80	160	1055	3,670.49	1	0.062	–	–	1	0.062

TABLE 7.3 Results of the Construction Heuristic and the VLSN Search Algorithm

		Construction Heuristic		VLSN Algorithm	
# of Weapons	# of Targets	Optimality Gap (%)	Time (sec)	Optimality Gap (%)	Time (sec)
10	5	0	<0.001	0	<0.001
10	10	0	<0.001	0	<0.001
10	20	0	<0.001	0	<0.001
20	10	0	<0.001	0	<0.001
20	20	0	<0.001	0	<0.001
20	40	0	0.015	0	0.015
20	80	0	0.015	0	0.031
40	10	1.79	0.015	0	0.031
40	20	0.33	0.015	0	0.015
40	40	0	0.015	0	0.015
40	80	0	0.031	0	0.078
40	120	0	0.062	0	0.109
80	20	2.33	0.109	0	0.156
80	40	0.10	0.062	0	0.109
80	80	0.0003	0.093	0.0003	0.156
80	160	0	0.172	0	0.219
80	320	0	0.390	0	0.625
100	50	0.79	0.120	0.0015	0.437
100	100	0.001	0.187	0.0009	0.250
100	200	0	0.375	0	0.609
200	100	0.01	0.656	0.0059	0.828
200	200	0.001	0.921	0.0008	1.109
200	400	0	1.953	0	2.516

80 weapons, VLSN algorithm enumerates neighborhood of size $80^5 \approx 3$ billion, which is very large. Third, our construction heuristic generates excellent starting solution for neighborhood search, which leads to a higher quality local optimal solution. The net effect of these three factors is that we get solutions that are very close to the optimal solution.

We observe that the construction heuristic obtained optimal solutions for over half of the instances and for the remaining instances the VLSN search algorithm converted them into optimal or almost optimal solutions. The computational times taken by these algorithms are also very low and even fairly large instances are solved within three seconds. It appears that the VLSN search algorithm is ideally suited to solve the WTA problem.

7.6 Conclusions

In this chapter, we consider the weapon target assignment problem which is considered to be one of the classical operations research problems that has been extensively studied in the literature but still has remained unsolved. Indeed, this problem is considered to be the holy grail of defense-related operations research. Though weapon target assignment problem is a nonlinear integer programming problem, we use its special structure to develop LP, MIP, network flow, and combinatorial lower bounding schemes. Using these lower bounding schemes in branch and bound algorithms gives us effective exact algorithms to solve the WTA problem. Our VLSN search algorithm also gives highly impressive results and gives either optimal or almost optimal solutions for all instances to which it is applied. To summarize, we can now state that the WTA problem is a well-solved problem and its large-scale instances can also be solved in real-time.

Acknowledgements

We gratefully acknowledge the insightful comments and suggestions of the two referees which helped us improve the presentation and clarify some results. We also wish to thank Mr Robert Murphey at the Air Force Research Laboratory, Eglin Air Force Base for bringing this problem to our notice and for helpful discussions. The research of the first three authors was partially supported by the NSF grant DMI-0217359. The fourth author gratefully acknowledges partial support from NSF grant DMI-0217123.

References

Ahuja, R.K., O. Ergun, J.B. Orlin, and A.P. Punnen. 2002. A survey of very large-scale neighborhood search techniques. *Discrete Applied Mathematics* 123, 75–102.

Ahuja, R.K., T.L. Magnanti, and J.B. Orlin. 1993. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, NJ.

Ahuja, R.K., J.B. Orlin, and D. Sharma. 2001. Multi-exchange neighborhood search algorithms for the capacitated minimum spanning tree problem. *Mathematical Programming* 91, 71–97.

Ahuja, R.K., J.B. Orlin, and D. Sharma. 2003. A composite very large-scale neighborhood structure for the capacitated minimum spanning tree problem. *Operations Research Letters* 31, 185–194.

Braford, J.C. 1961. Determination of optimal assignment of a weapon system to several targets. AER-EITM-9, Vought Aeronautics, Dallas, TX.

Castanon, D.A. 1987. Advanced weapon-target assignment algorithm. Quarterly Report # TR-337, ALPHA TECH, Inc., Burlington, MA.

Chang, S.C., R.M. James, and J.J. Shaw. 1987. Assignment algorithm for kinetic energy weapons in boost defense. *Proceedings of the IEEE 26th Conference on Decision and Control*, Los Angeles, CA.

Day, R.H. 1966. Allocating weapons to target complexes by means of nonlinear programming. *Operations Research* 14, 992–1013.

DenBroeder, G.G., Jr., R.E. Ellison, and L. Emerling. 1958. On optimum target assignments. *Operations Research* 7**,** 322–326.

Eckler, A.R., and S.A. Burr. 1972. *Mathematical Models of Target Coverage and Missile Allocation*. Military Operations Research Society, Alexandria, VA.

Fredman, M.L., and R.E. Tarjan. 1984. Fibonacci heaps and their uses in improved network optimization algorithms. *Proceedings of the 25th Annual IEEE Symposium on Foundations of Computer Science*, 338–346. Full paper in *Journal of ACM* 34 (1987), 596–615.

Gendreau, M., Guertin, F., Potvin, J.Y., and R. Seguin. 2006. Neighborhood search heuristics for a dynamic vehicle dispatching problem with pick-ups and deliveries. *Transportation Research Part C* 14, 157–174.

Glover, F. 1992. Ejection chains, reference structures, and alternating path algorithms for the traveling salesman problem, Research report, University of Colorado-Boulder, Graduate School of Business. (A short version appeared in Discrete Appl. Math. 65 (1996) 223–253)

Green D.J., J.T. Moore, and J.J. Borsi. 1997. An integer solution heuristic for the arsenal exchange model (AEM). *Military Operations Research Society* 3 (2), 5–15.

Katter, J.D. 1986. A solution of the multi-weapon, multi-target assignment problem. Working paper 26957, MITRE.

Lloyd, S.P., and H.S. Witsenhausen. 1986. Weapons allocation is NP-complete. *Proceedings of the 1986 Summer Conference on Simulation*, Reno, NV.

Maltin, S.M. 1970. A review of the literature on the missile-allocation problem. *Operations Research* 18, 334–373.

Manne, A.S. 1958. A target-assignment problem. *Operations Research* 6, 346–351.

Metler, W.A., and F.L. Preston. 1990. A suite of weapon assignment algorithms for a SDI mid-course battle manager. NRL Memorandum Report 671, Naval Research Laboratory, Washington, DC.

Murphey, R.A. 1999. Target-based weapon target assignment problems. In *Nonlinear Assignment Problems: Algorithms and Applications*. P.M. Pardalos and L.S. Pitsoulis (Eds.), Norwell, MA: Kluwer Academic Publishers, 39–53.

Murty, K.G. 1976. *Linear and Combinatorial Optimization*. Somerset, NJ: John Wiley.

Orlin, D. 1987. Optimal weapons allocation against layered defenses. *Naval Research Logistics* 34, 605–616.

Rego, C., and C. Roucairol. 1996. A parallel tabu search algorithm using ejection chains for the vehicle routing problem. In *Meta-Heuristics: Theory and Applications.* I.H. Osman, J.P. Kelly (Eds.), Dordrecht: Kluwer Academic Publishers.

Thompson, P.M., and J.B. Orlin. 1989. The theory of cyclic transfers. MIT Operations Research Center Report 200-89, Cambridge, MA.

Thompson, P.M., and H.N. Psaraftis. 1993. Cyclic transfer algorithms for multi-vehicle routing and scheduling problems. *Operations Research* 41**,** 935–946.

Wacholder, E. 1989. A neural network-based optimization algorithm for the static weapon-target assignment problem. *ORSA Journal on Computing* 4, 232–246.

Xu, J., and J. Kelley. 1996. A network-flow based tabu seach heuristic for the vehicle routing problem. *Transportation Science,* 30, 379–393.

8

Optimization Model for Military Budget Allocation and Capital Rationing

Adedeji B. Badiru
Air Force Institute of Technology

Charles H. Aikens
University of Tennessee

8.1 Introduction

Protection of national interests is accomplished over several capabilities. These capabilities must all be funded at appropriate levels and on a relative basis in order to achieve desired outcomes. In an era of diminishing budgetary resources, strategic allocation of limited resources must be exercised. It has been shown that heuristic approaches to budget allocation do not work effectively for complex decision problems. Military operations constitute such a complex decision problem (Alberts and Hayes 2007). As warfare becomes more network-centric (Moffat 2006), strategies must be developed to ensure that each element of the composite network receives mission-appropriate resource allocation. This is accomplished through optimization of budget allocation and capital rationing. This chapter presents an example of a linear programming model developed for budget allocation by the Commander of Navy Installations (CNI). The modeling approach can be adapted to other military or commercial applications.

8.2 Command and Control Domains

Initiation and execution of military missions are accomplished within the domains of "command and control (C2)" processes (Alberts et al. 2007). Command and control is the exercise of authority and direction by a properly designated commander over assigned and attached forces in the accomplishment of the mission. Command and control functions are performed through an arrangement of personnel, equipment, communications, facilities, and procedures employed by a commander in planning, directing, coordinating, and controlling forces and operations in the accomplishment of the mission. Put simply, C2 ensures that information and other resources get to those performing specific mission-critical functions at appropriate and effective times.

C2 is operationally extended to "command, control, communications, computing, and intelligence (C4I)", which represents systems that are essential to national security. There are several historical accounts of how C4I influenced the outcome of an engagement or an entire conflict. Thus, C4I is an essential capability that must be addressed in the pursuit of national security. But it must compete for funding with other crucial capabilities desired by the military.

8.3 Budgeting Across Capabilities

Strategic budget allocation involves sharing limited resources between several operational objectives in a strategic cost management challenge. Considering each budget allocation as an investment, the valuation (Damodaran 1996) of such investment represents a measure of outcome relative to the input of budgetary resources in the presence of risks. Even in a nonprofit environment, investment analysis and management (Jones 2000) are essential for ensuring operational effectiveness. Budget and investment analysis can serve any of the following purposes:

- A plan for resources expenditure
- A project selection criterion
- A projection of project policy
- A basis for project control
- A performance measure
- A standardization of resource allocation
- An incentive for process improvement

Top-down budgeting involves collecting data from upper level sources such as top and middle managers. The figures supplied by the managers may come from their personal judgment, past experience, or past data on similar project activities. The cost estimates are passed to lower level managers, who then break the estimates down into specific work components within the project. These estimates may, in turn, be given to line managers, supervisors, and lead workers to continue the process until individual activity costs are obtained. Top management provides the global budget, while the functional level worker provides specific budget requirements for project items.

In bottom-up budgeting, elemental activities, and their schedules, descriptions, and labor skill requirements are used to construct detailed budget requests. Line workers familiar with specific activities are requested to provide cost estimates. Estimates are made for each activity in terms of labor time, materials, and machine time. The estimates are then converted to an appropriate cost basis. The dollar estimates are combined into composite budgets at each successive level up the budgeting hierarchy. If estimate discrepancies develop, they can be resolved through the intervention of senior management, middle management, functional managers, project manager, accountants, or standard cost consultants.

8.4 General Formulation of Budget Allocation Problem

A general formulation for capital rationing involves selecting a combination of projects that will optimize the return on investment or maximize system effectiveness. A general formulation of the capital budgeting (Badiru and Omitaomu 2007) problem is presented:

$$\text{Maximize } z = \sum_{i=1}^{n} v_i x_i$$

$$\text{Subject to } \sum_{i=1}^{n} c_i x_i \leq B$$

$$x_i = 0, 1; \quad i = 1, \ldots, n$$

where n = number of projects; v_i = measure of performance for project i (e.g. present value); c_i = cost of project i; x_i = indicator variable for project i; B = budget availability level.

A solution of the above model will indicate what projects should be selected in combination with other projects. The example that follows illustrates a capital rationing problem. Planning a portfolio of projects is essential in resource-limited projects. The capital-rationing formulation that follows demonstrates how to determine the optimal combination of project investments (or budget allocations) so as to maximize total return on investment or total system effectiveness. Suppose a project analyst is given N projects, $X_1, X_2, X_3, \ldots, X_N$, with the requirement to determine the level of investment in each project so that total investment return is maximized subject to a specified limit on available budget. We assume that the projects are not mutually exclusive. The investment in each project starts at a base level b_i ($i = 1, 2, \ldots, N$) and increases by variable increments k_{ij} ($j = 1, 2, 3, \ldots, K_i$), where K_i is the number of increments used for project i. Consequently, the level of investment in project X_i is defined as follows:

$$x_i = b_i + \sum_{j=1}^{K_i} k_{ij}$$

where

$$x_i \geq 0, \forall i$$

For most cases, the base investment will be zero. In those cases, we will have $b_i = 0$. In the modeling procedure used for this problem, we have:

$$X_i = \begin{cases} 1 & \text{if the investment in project } i \text{ is greater than zero} \\ 0 & \text{otherwise} \end{cases}$$

and

$$Y_{ij} = \begin{cases} 1 & \text{if } j\text{th increment of alternative } i \text{ is used} \\ 0 & \text{otherwise} \end{cases}$$

The variable x_i is the actual level of investment in project i, while X_i is an indicator variable indicating whether or not project i is one of the projects selected for investment. Similarly, k_{ij} is the actual magnitude of the jth increment while Y_{ij} is an indicator variable that indicates whether or not the jth increment is used for project i. The maximum possible investment in each project is defined as M_i, such that:

$$b_i \le x_i \le M_i.$$

There is a specified limit, B, on the total budget available to invest, such that

$$\sum_i x_i \le B$$

There is a known relationship between the level of investment, x_i, in each project and the expected return, $R(x_i)$. This relationship will be referred to as the *utility function*, $f(.)$, for the project. The utility function may be developed through historical data, regression analysis, and forecasting models. For a given project, the utility function is used to determine the expected return, $R(x_i)$, for a specified level of investment in that project. That is,

$$R(x_i) = f(x_i)$$
$$= \sum_{j=1}^{K_i} r_{ij} Y_{ij}$$

where r_{ij} is the incremental return obtained when the investment in project i is increased by k_{ij}. If the incremental return decreases as the level of investment increases, the utility function will be concave. In that case, we will have the following relationship:

$$r_{ij} \ge r_{ij+1}$$
$$r_{ij} - r_{ij+1} \ge 0$$

Thus,

$$Y_{ij} \ge Y_{ij+1}$$
$$Y_{ij} - Y_{ij+1} \ge 0$$

so that only the first n increments ($j=1, 2, \ldots, n$) that produce the highest returns are used for project i. Figure 8.1 shows an example of a concave investment utility function.

If the incremental returns do not define a concave function, $f(x_i)$, then one has to introduce the inequality constraints presented above into the optimization model. Otherwise, the inequality constraints may be left out of the model, since the first inequality, $Y_{ij} \ge Y_{ij+1}$, is always implicitly satisfied for concave functions. The objective is to maximize the total return. That is,

$$\text{Maximize } Z = \sum_i \sum_j r_{ij} Y_{ij}$$

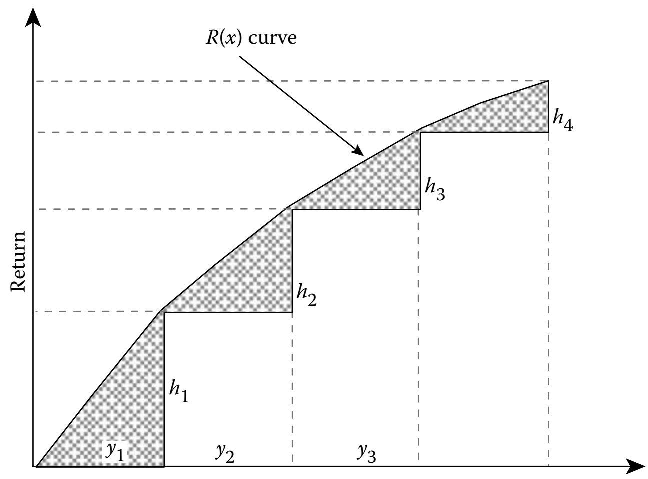


FIGURE 8.1 Utility curve for investment yield.

Subject to the following constraints:

$$x_i = b_i + \sum_j k_{ij} Y_{ij} \quad \forall i$$

$$b_i \le x_i \le M_i \quad \forall i$$

$$Y_{ij} \ge Y_{ij+1} \quad \forall i, j$$

$$\sum_i x_i \le B$$

$$x_i \ge 0 \quad \forall i$$

$$Y_{ij} = 0 \text{ or } 1 \quad \forall i, j$$

8.5 Application to Military Budget Allocation

Optimization modeling is of interest for military budget allocation and capital rationing. The sections that follow are based on an optimization modeling project that the authors did for the US Navy in early 2000s. With increasing demands on the US war-fighting capabilities, it has become more important to develop strategic cost management initiatives for the military. Cost is the driving factor in many recent decisions to deploy or not deploy new or enhanced technology-based war-fighting equipment. Noted examples include MRAP (mine-resistant ambush-protected) vehicle (for US Marines) and C-5A Galaxy transport plane (for US Air Force).

8.6 Problem Background

The US Navy is committed to protecting personnel, assets, and mission capability at its installations around the world. As part of this initiative, the Navy is evaluating and implementing antiterrorism (AT) capabilities that will increase the Navy's ability to deter, detect/inspect, and respond to terrorist threats. While many capabilities already exist and others are being developed, the Navy must make decisions on how to allocate resources in a resource-constrained environment to best manage the risks associated with security threats. The Commander, Navy Installations (CNI) is using a risk-based approach to allocate limited resources and would like to augment previous and ongoing risk analysis work by using operations research (OR) techniques to employ a rigorous approach to resource allocation and optimization. In the proposed formulation, the treatment of risk is analogous to the treatment of investment yield; in the sense that investment yield may represent a measure of risk mitigation.

CNI is currently considering 21 specific AT capabilities for implementation at installations around the world. Implementation of the capabilities requires funding from one or more of the following four accounts over a six-year period from fiscal year (FY) 2006 through FY 2011:

1. Research development testing and evaluation (RDTE)
2. Operations and maintenance (OMN)
3. Other procurement (OPN)
4. Military construction (MCON)

CNI has teamed with several consulting firms, one of which developed a prototype linear program (LP) model (Badiru and Aikens 2005) that demonstrated that LP can be used to address this problem. The model seeks to optimize resource allocation among the 21 AT capabilities. The sections that follow represent a summary of the results of the analysis performed by Badiru and Aikens (2005).

8.7 Problem Statement

Optimize the allocation of four types of funding between 21 AT capabilities over multiple years to achieve maximum risk reduction (or yield optimization).

8.8 Model Scenarios

- Given the large number of constraints and rules that are anticipated, a full model may not have an optimized solution. To the extent possible, the model should include constraint equations formulated in a way to help prevent this situation (e.g. expressed as inequalities or ranges around a particular input value).
- CNI will eventually set a constraint for each of the four types of funding for each FY between FY06 and FY11 (six years by four types of funding, resulting in 24 constraint equations involving funding), but it is likely that this effort will be iterative based on "what if" type modeling.
- CNI will probably want to optimize funding types (i.e. determine if changing from one type of funding to another within a FY will impact total risk reduction), but it is likely that this will be done based on a review of model results, particularly marginal benefits and sensitivity analysis data (e.g. the model will permit manual adjustments to rerun the model rather than determining an optimum funding strategy).
- CNI will probably want to perform sensitivity analyses on model constraints such as overall funding level, changing types of money, and moving funding from one year to another year (presumably, they will only choose to model changes in funding from one year to the next if the time dimension is enabled).
- CNI will probably want to perform some sort of sensitivity analysis on the risk-related data that are used in the objective function, possibly via simulation.

8.9 Strategy for Good Modeling Approach

- Construct an OR-LP model that will maximize risk reduction subject to a set of constraints.
- Run, verify, and validate the model.
- Develop and test various types of constraints that are needed to produce logical model behavior.

- Understand data needs and the relative importance of data that are not currently available.
- Investigate the following performance metrics:
  - Sensitivity.
    - How sensitive is Z (the objective variable) to changes in data or constraints?
  - Marginal benefits.
    - What risk reduction can be achieved by investing an additional dollar in a capability?
  - Feasibility.
    - Is there a solution that meets all of the model constraints?
- The model uses LINGO, a Windows-based linear and nonlinear programming model solver.
- LINGO can import data from an Excel workbook.

8.10 The Objective Function

The objective is to maximize the reduction in risk by allocating resources to one or more capabilities based on the following expression

$$Z = \sum_{i=1}^{n} C_i.\Phi_i.Y_i$$

where Z, the objective variable, is the sum of the products of the total reduction in risk that is achievable for each capability times the fraction of the risk reduction that is achieved by a given level of funding (i.e. the utility of the spend). C_i represents the reduction in risk resulting from full implementation of activity i (from the risk-based investment strategy report). Φ_i is a set of utility functions for each capability that maps funding level to the fraction of risk that will be achieved for a particular funding level (ranging from 0 to 100% of the requested budget and 0 to 100% of C_i). Y_i is a binary variable that indicates if a capability is "turned on" (binary variables are mathematically useful for expressing constraints/rules such as interdependency between capabilities).

8.11 Model Representation for the Prototype Run

As a preliminary step, the problem dimensions were scoped down from 21 capabilities to six and the utility functions were assumed to be linear and convex. As a future model enhancement, the convexity assumption makes it possible to handle nonlinearities through a piecewise linearization of each function thereby expanding the number of variables accordingly. The pilot model is formulated as follows:

$$\phi(x_{ij}) = \alpha_{ij} + \beta_{ij} x_{ij} = \text{the linear approximation to the utility function for capability } i \text{ and funding source } j.$$

where x_{ij} = proportion of full funding budget for capability i from funding source j that is actually allocated to capability i; $x_{ij} = 1$ represents full funding; and c_i = risk reduction coefficient for capability i.

$$y_i = \begin{cases} 1 & \text{if capability } i \text{ is deployed} \\ 0 & \text{otherwise} \end{cases}$$

$$0 \le x_{ij} \le 1; \quad i = 1, 2, \ldots, 6 \quad j = 1, 2, \ldots, 4$$

$$\text{Objective:} \quad \text{Max} \quad z = \sum_i \left(c_i \big/ n_i \right) y_i \sum_j \left(\alpha_{ij} + \beta_{ij} x_{ij} \right)$$

n_i = the number of funding sources required for capability i. $0 \le n_i \le 4$, i = 1, 2, …, 6.

Budget constraints:

1. $\sum_i r_{ij} x_{ij} \le B_j \qquad j = 1, 2, ..., 4$

Total amount ($) allocated to capability i from funding j cannot exceed the total budget for funding source j. r_{ij} = budget from funding source j needed to implement capability i at 100%.

Contingency constraints:

1. $y_{i't} - y_{i''t} \ge 0 \qquad i', i'' \in \{I\}$ = set of capabilities.
   If capability i' is implemented, capability i'' cannot be implemented.
2. $y_{i't} - y_{i''t} = 0 \qquad i', i'' \in \{I\}$ = set of capabilities.
   If capability i' is implemented, capability i'' must also be implemented.
3. $y_i - x_{ij} \le 0.999 \qquad j' \in \{R_i\}$ = set of resource requirements for capability i.
   If capability i is implemented, all resource requirements must be deployed.

Deployment status:

1. $x_{ij} - y_j \le 0 \qquad i = 1, 2, \ldots, 6 \qquad j = 1, 2, \ldots, 4$

Dollars can only be allocated to those capabilities that have been implemented.

Balance equations:

1. $\alpha_{ij} + \beta_{ij} x_{ij} = \alpha_{ik} + \beta_{ij} x_{ik} \qquad \text{i} = 1, 2, \ldots, 6 \qquad j \ne k$

The benefits derived across utility functions for each capability should be equal to avoid wasting resources. Activating these constraints will likely lead to infeasibility.

Minimum funding:

1. $x_{ij} - \rho_{ij} y_i \ge 0$

where ρ_{ij} is the minimum set funding level from source j and capability i. Capability will only be implemented if funding is above a minimum level.

Bounding:

1. $L_{ij} \le x_{ij} \le 1 \qquad i = 1, 2, \ldots, 6 \qquad j = 1, 2, \ldots, 4$

Allocations must lie between some lower bound and 1. The lower bound may be 0.

2. $y_i \in \{0,1\} \qquad i = 1, 2, \ldots, 6$

Represents implementation decisions.

8.12 Utility Function Modeling Approach

Utility functions (Figure 8.2) must be defined for each capability and each type of resource (i.e. funding)

- A utility function maps the amount of resources applied to a particular capability to the utility derived.
- Utility functions are rarely linear from (0,0) to (1,1).

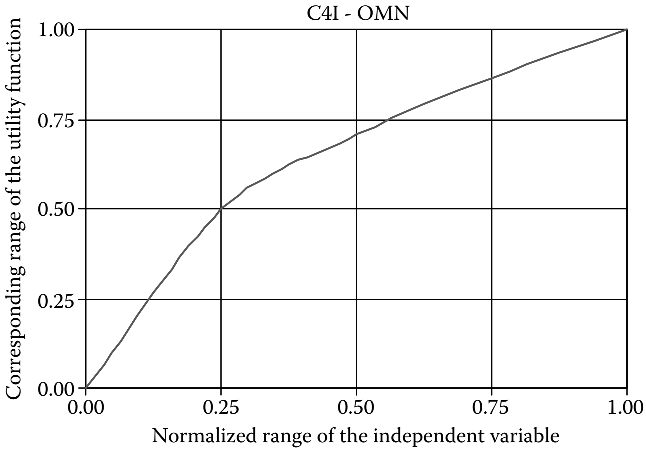


FIGURE 8.2 Example of utility function C4I-OMN.

Simplifying assumptions:

- Initial problem limited to six capabilities
  1. Decision making system
  2. Area wide alert/notification
  3. Operations center
  4. Security surveillance
  5. CBRNE response (Chemical, Biological, Radiation, Nuclear, Explosives)
  6. C4I (Command, Control, Communications, Computing, and Intelligence)
- Resource constraints are based on total funding (not by FY at this point)
- Linear utility and constraint functions

Data used in the model:

- Risk coefficients for the objective function (C_i), see Table 8.1
- Costs
  - Costs for full implementation of each activity from tentative budget data provided on February 12, 2004 (with hypothetical MCON costs)
- Constraints:
  - Cost constraints (hypothetical values were used to test the model)
  - Interdependencies (hypothetical values were used, but interdependencies are currently being evaluated as part of this project)
  - Other logical constraints

TABLE 8.1 Risk Reduction Associated with Providing 100% of a Capability (the c_i's)

Capability	Risk
Decision making system	87
Area wide alert/notification	70
Operations center	90
Security surveillance	42
CBRN response	79
C4I	90

- Utility functions for each capability.
    - One utility function for each type of money, up to four functions per capability (hypothetical utility functions were used, but functions could be developed using a risk analysis tool that estimates marginal risk reduction).

Financial constraints:

- For model demonstration purposes, the following program funding levels were established (the B_j's)
    - MCON, \$35.00
    - RDT&E, \$20.00
    - OPN, \$300.00
    - OMN, \$60.00

Interdependencies:

- The interdependencies were recognized (but not explicitly dealt with) in the original risk-based investment strategy resource allocation model.
    - Using constraint equations, dependencies can be established in forms such as:
        - If capability A is implemented, then capability B must also be implemented
        - If capability C is not implemented, then capability D must not be implemented
        - Either capability E or capability F are implemented, but not both
        - Other interdependencies between capabilities that are rules driven
- Interdependencies are a common LP constraint

Other constraints:

- Nonnegativity
- Minimum funding (e.g. avoid regions of large error between "actual"and "linear" utility functions
- Balance equations (Figure 8.3)

Data needed for each capability:

- Description (from Navy readiness documents)
- Cost (from tentative data provided by CNI)

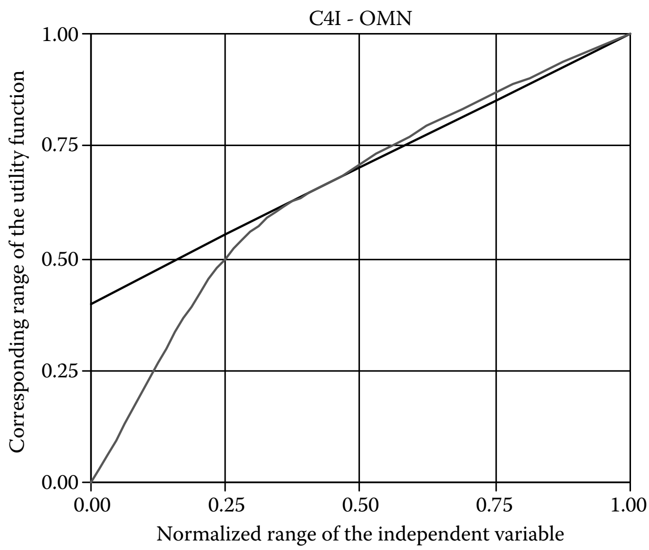


FIGURE 8.3 Plot of balance equations for C4I-OMN.

- Utility functions

Capability 1: Decision making system.

- A method to determine tactics, techniques, procedures, and rules of engagement necessary to alleviate the threat. Figure 8.4 illustrates the utility functions for capability 1
- Cost
  - MCON, \$0.00
  - RDT&E, \$9.00
  - OPN, \$149.30
  - OMN, \$12.90

$$f(x_{1,\text{RDT\&E}}) = f(x_{12}) \cong -3 + 4x_{12}$$

$$f(x_{1,\text{OPN}}) = f(x_{13}) \cong 0.4 + 0.6x_{13}$$

$$f(x_{1,\text{OMN}}) = f(x_{14}) \cong -0.33 + 1.33x_{14}$$

Capability 2: Area-wide alert/notification.

- Methods to update and disseminate information concerning a terrorist threat or danger, including procedures and techniques necessary to counter the threat (may include civil distribution systems). Figure 8.5 shows the alert and notification utility functions for capability 2.
- Cost
  - MCON, \$8.00
  - RDT&E, \$0.00
  - OPN, \$74.30
  - OMN, \$5.50

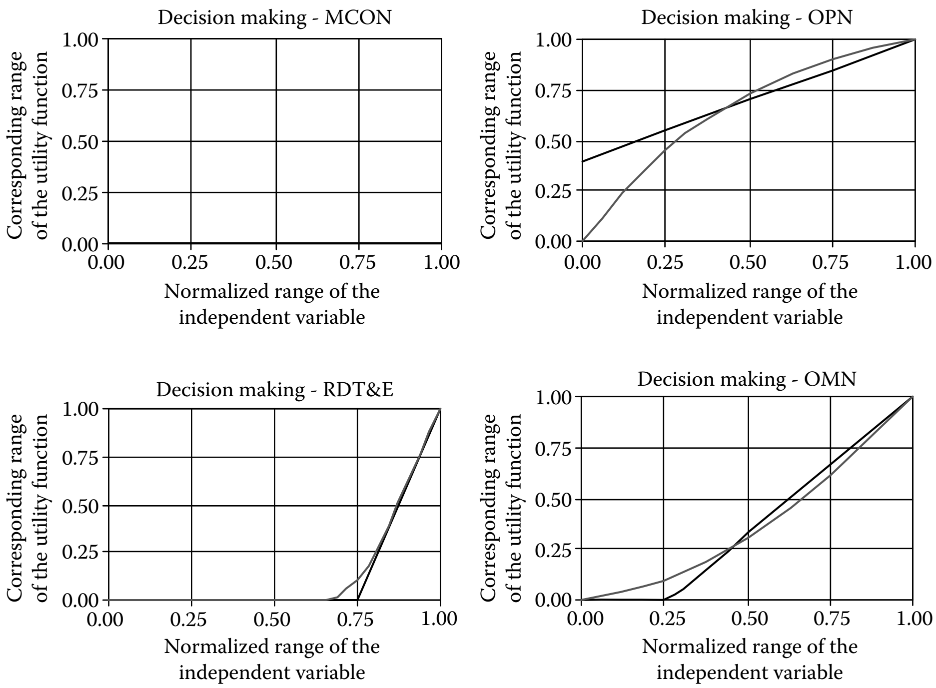


FIGURE 8.4 Decision making system utility functions for capability 1.

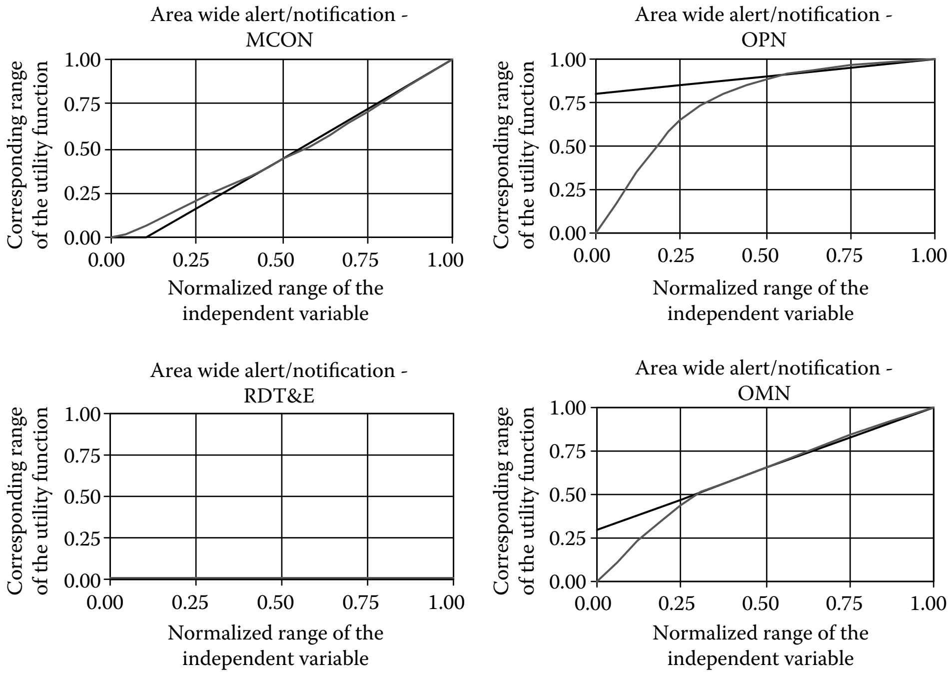


FIGURE 8.5 Area-wide alert and notification utility functions.

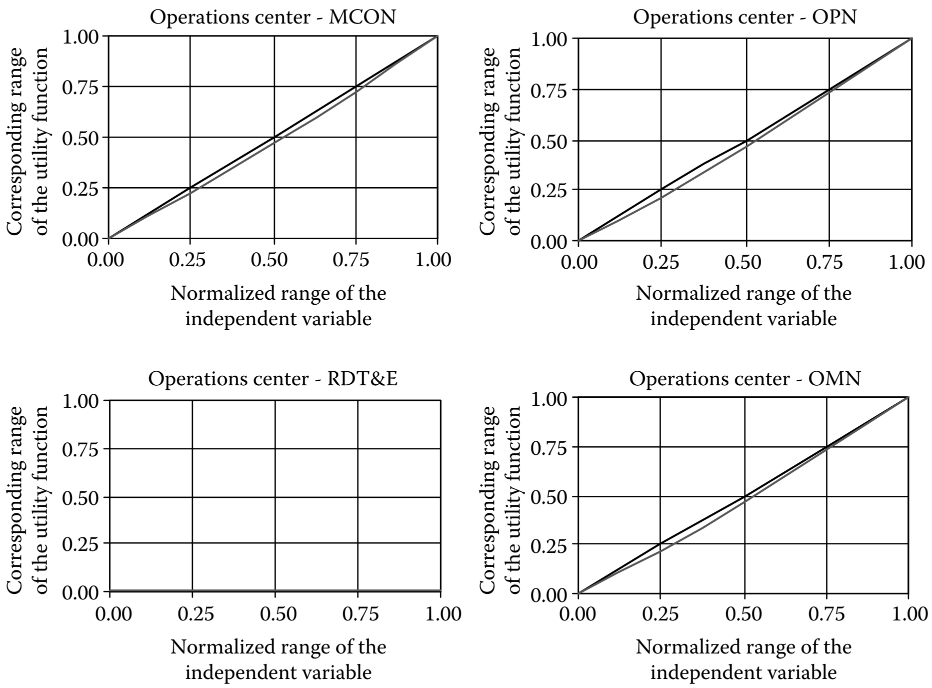


FIGURE 8.6 Operations center utility functions.

$$f(x_{2,\text{MCON}}) = f(x_{21}) \cong -0.11 + 1.1x_{21}$$

$$f(x_{2,\text{OPN}}) = f(x_{23}) \cong 0.8 + 0.2x_{23}$$

$$f(x_{2,\text{OMN}}) = f(x_{24}) \cong 0.3 + 0.7x_{24}$$

Capability 3: Operations center.

- Locations containing decision-making capability for command, control, communication, and computer systems. Figure 8.6 shows the utility functions for capability 3.
- Cost
  - MCON, $94.00
  - RDT&E, $0.00
  - OPN, $80.60
  - OMN, $2.50

$$f(x_{3,\mathrm{MCON}}) = f(x_{31}) = x_{31}$$

$$f(x_{3,\mathrm{OPN}}) = f(x_{33}) = x_{33}$$

$$f(x_{3,\mathrm{OMN}}) = f(x_{34}) = x_{34}$$

Capability 4: Security surveillance (nonSWIMS).

- Detecting and identifying unauthorized movement of personnel and material to installations, piers, harbors, ships, and incident sites from land. Figure 8.7 shows the utility functions for capability 4.
- Cost
  - MCON, $7.00
  - RDT&E, $0.00
  - OPN, $170.60
  - OMN, $49.20

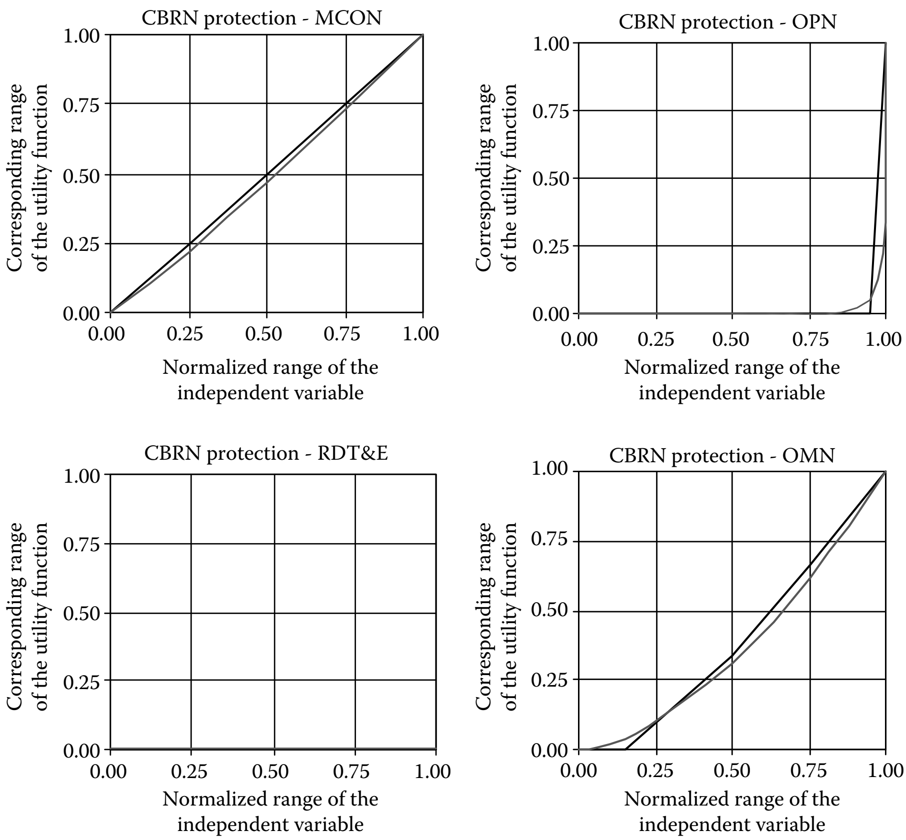


FIGURE 8.7 Security surveillance utility functions.

$$f(x_{4,\text{MCON}}) = f(x_{41}) \cong x_{41}$$

$$f(x_{4,\text{OPN}}) = f(x_{43}) \cong -19 + 20x_{43}$$

$$f(x_{4,\text{OMN}}) = f(x_{44}) \cong -0.18 + 1.18x_{44}$$

Capability 5: CBRNE response.

- Includes
  - decontaminating personnel and casualties, and disarming and disposing of explosive devices;
  - rescuing individuals, extinguishing fires, and securing utilities at an incident site;
  - collecting and preserving evidence after a CBRNE (Chemical, Biological, Radiation, Nuclear, Explosive) incident. Figure 8.8 presents the utility functions for capability 5.
- Cost
  - MCON, $0.00
  - RDT&E, $12.00
  - OPN, $8.70
  - OMN, $57.50

$$f(x_{5,\text{RDT\&E}}) = f(x_{52}) \cong -1.5 + 2.5x_{52}$$

$$f(x_{5,\text{OPN}}) = f(x_{53}) \cong -0.67 + 1.67x_{53}$$

$$f(x_{5,\text{OMN}}) = f(x_{54}) \cong 0.9 + 0.1x_{54}$$

Capability 6: C4I.

- Includes
  - tasks within command, control, communications, and computer systems associated with installation AT/FP operations;

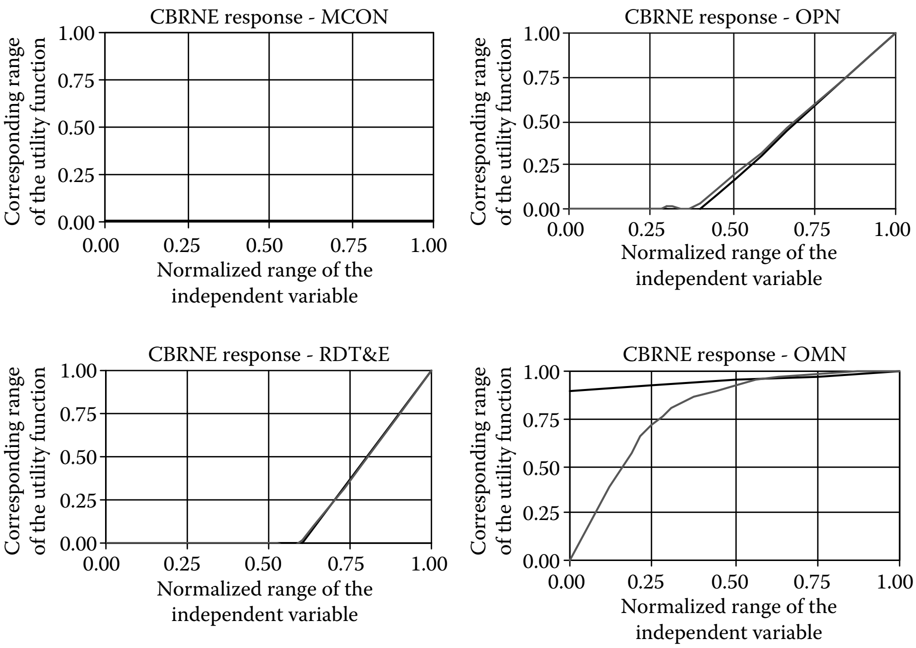


FIGURE 8.8 Utility functions for CBRNE response (capability 5).

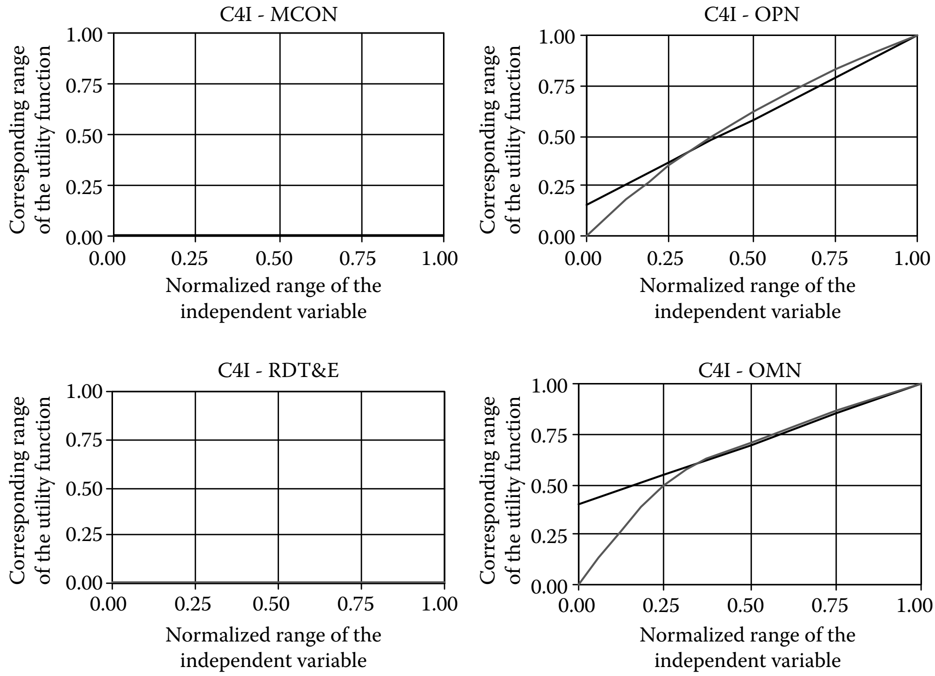


FIGURE 8.9 Utility functions for capability 6 (C4I).

  - collecting and evaluating hazard information, including tracking, mapping, and marking contamination movement, and contamination impact on the environment and infrastructure. Figure 8.9 presents the utility functions for capability 6 (C4I).
- Cost
  - MCON, \$0.00
  - RDT&E, \$0.00
  - OPN, \$392.30
  - OMN, \$43.40

$$f(x_{6,\text{OPN}}) = f(x_{63}) \cong 0.15 + 0.85x_{63}$$

$$f(x_{6,\text{OMN}}) = f(x_{64}) \cong 0.4 + 0.6x_{64}$$

8.13 The Prototype Model

$$\text{Objective:} \quad \text{Max} \quad z = \left(87/3\right) y_1((-3+4x_{12}) + (0.4+0.6x_{13}) + (-0.33+1.33x_{14}))$$

$$+\left(70/3\right) y_2((-0.11+1.1x_{21}) + (0.8+0.2x_{23}) + (0.3+0.7x_{24}))$$

$$+\left(90/3\right) y_3(x_{31}+x_{33}+x_{34}) + \left(42/3\right) y_4(x_{41} + (-19+20x_{43}) + (-0.18+1.18x_{44}))$$

$$+\left(79/3\right) y_5((-1.5+2.5x_{52}) + (-0.67+1.67x_{53}) + (0.9+0.1x_{54}))$$

$$+\left(90/2\right) y_6((0.15+0.85x_{63}) + (0.4+0.6x_{64}))$$

$$\text{Max}\quad z=-84.97y_1+23.1y_2-268.52y_4-33.44y_5+24.75y_6$$
$$+116y_1x_{12}+17.4y_1x_{13}+38.57y_1x_{14}+25.67y_2x_{21}+4.67y_2x_{23}+16.33y_2x_{24}$$
$$+30y_3x_{31}+30y_3x_{33}+30y_3x_{34}+14y_4x_{41}+280y_4x_{43}+16.52y_4x_{44}+65.83y_5x_{52}$$
$$+43.98y_5x_{53}+2.63y_5x_{54}+38.25y_6x_{63}+27y_6x_{64}$$

Subject to:

1. $8x_{21}+94x_{31}+7x_{41}\leq 35$
2. $9x_{12}+12x_{52}\leq 20$
3. $149.3x_{13}+74.3x_{23}+80.6x_{33}+170.6x_{43}+8.7x_{53}+392.3x_{63}\leq 300$
4. $12.9x_{14}+5.5x_{24}+2.5x_{34}+49.2x_{44}+57.5x_{54}+43.4x_{64}\leq 60$
5. $x_{12}-y_1\leq 0$
6. $x_{13}-y_1\leq 0$
7. $x_{14}-y_1\leq 0$
8. $x_{21}-y_2\leq 0$
9. $x_{23}-y_2\leq 0$
10. $x_{24}-y_2\leq 0$
11. $x_{31}-y_3\leq 0$
12. $x_{33}-y_3\leq 0$
13. $x_{34}-y_3\leq 0$
14. $x_{41}-y_4\leq 0$
15. $x_{42}-y_4\leq 0$
16. $x_{43}-y_4\leq 0$
17. $x_{44}-y_4\leq 0$
18. $x_{52}-y_5\leq 0$
19. $x_{53}-y_5\leq 0$
20. $x_{54}-y_5\leq 0$
21. $x_{63}-y_6\leq 0$
22. $x_{64}-y_6\leq 0$
23. $y_1-x_{12}\leq 0.999$
24. $y_1-x_{13}\leq 0.999$
25. $y_1-x_{14}\leq 0.999$
26. $y_2-x_{21}\leq 0.999$
27. $y_2-x_{23}\leq 0.999$
28. $y_2-x_{24}\leq 0.999$
29. $y_3-x_{31}\leq 0.999$
30. $y_3-x_{33}\leq 0.999$
31. $y_3-x_{34}\leq 0.999$
32. $y_4-x_{41}\leq 0.999$
33. $y_4-x_{42}\leq 0.999$
34. $y_4-x_{43}\leq 0.999$
35. $y_4-x_{44}\leq 0.999$
36. $y_5-x_{52}\leq 0.999$
37. $y_5-x_{53}\leq 0.999$
38. $y_5-x_{54}\leq 0.999$
39. $y_6-x_{63}\leq 0.999$
40. $y_6-x_{64}\leq 0.999$
41. $y_1-y_4=0$
42. $y_2-y_1=0$
43. $y_2-y_3=0$

44. $y_5 - y_1 = 0$
45. $y_6 - y_1 = 0$
46. $x_{13} - 0.3y_1 \geq 0$
47. $x_{14} - 0.4y_1 \geq 0$
48. $x_{23} - 0.5y_2 \geq 0$
49. $x_{24} - 0.25y_2 \geq 0$
50. $x_{52} - 0.25y_5 \geq 0$
51. $x_{44} \geq 0.1$
52. $x_{42} \geq 0.25$

Numerical results (Solution 1 below):

$y_1 = 1, y_2 = 1, y_3 = 1, y_4 = 1, y_5 = 1, y_6 = 1$
$x_{12} = 1, x_{13} = 0.3, x_{14} = 0.4$
$x_{21} = 1, x_{23} = 0.5, x_{24} = 0.25$
$x_{31} = 0.213, x_{33} = 0.001, x_{34} = 0.742$
$x_{41} = 1, x_{42} = 0.001, x_{43} = 0.652, x_{44} = 0.2$
$x_{52} = 0.917, x_{53} = 1, x_{54} = 0.5$
$x_{63} = 0.25, x_{64} = 0.3, z = 181.45$

Solution interpretation

Deploy all capabilities:

	MCON	RDT&E	OPN	OMN
Decision making	–	100%	30%	40%
Area wide alert	100%	–	50%	25%
Operations center	21.3%	–	–	74.2%
Security surveillance	100%	–	65.2%	20%
CBRN response	–	91.7%	100%	50%
C4I	–	–	25%	30%

This solution was obtained by setting a lower bound on RDT&E and OMN for security and surveillance at 0.25 and 0.1, respectively (constraints 51 and 52).

Local optimal solution found at iteration: 6

Objective value: 181.450.

Comment: Note how the model takes funding from some capabilities in order to fund others (e.g. OPN and OMN funding is distributed across numerous capabilities.

To illustrate how the model might respond to a relaxation of contingency constraints, constraint 44 was removed, no longer requiring the simultaneous funding of capabilities one and five. This produced the revised solution shown below (Solution 2) which favors not deploying capability four.

Deploy all capabilities except capability 5:

	MCON	RDT&E	OPN	OMN
Decision making	–	–	30%	40%
Area wide alert	100%	–	50%	25%
Operations center	21.3%	–	–	–
Security and surveillance	100%	25%	70.3%	20%
CBRN response	–	–	–	–
C4I	–	–	25%	30%

Objective function value: –14.584

8.14 Prototype Model Results

- LP can be used to optimize resource allocation, but only if
  - Risk is understood for each capability/task (i.e. output of initial RBDM investment strategy effort).
  - Utility functions are understood for each task and type of funding (possible output from at least one approach being considered for SYSCOM model).
- Even if there is considerable uncertainty in this information, the model can be used to better understand bottlenecks, constraints, and the effect of dependencies between capabilities/tasks.

8.15 Possible Modeling Enhancements

- Evaluate all 21 capabilities (or 64 tasks) based on:
  - Defendable utility functions
  - Improved understanding of interdependencies
  - Actual financial constraints
- Piecewise-linear (or nonlinear) utility functions
- Enhance the model to incorporate a time dimension
- Understand year-to-year risk for a particular investment strategy
- Define interdependencies
- Develop utility functions
- Expand the LP model:
  - Address all capabilities/tasks of interest
  - Add a time dimension
  - Improve the balance functions

8.16 Phase One Outputs

1. Development of a prototype model that demonstrates how OR could be used to approach the budget allocation problem. The illustrative (prototype) model was scoped down to only six of the 21 capability areas. More robust models can cover all 21 capability areas.
2. Compilation of a list of assumptions used to develop the model and constraint equations.
3. Collation of a list of questions and/or data that the Navy will eventually need to develop a fully functional model. The data needs identified in the initial phase include:
   - The form of the utility functions that map funding for a particular capability to reduction in risk and data needed to characterize the utility function for each capability.
   - A full suite of inter-capability dependencies that will allow modelers to prevent the model from overfunding capability A without adequately funding other capability(ies) that must be in place to obtain the risk reduction depicted by the utility function for capability A.
   - Funding constraints, as well as any other constraints that are not obvious to the modeling team (e.g. nonnegativity is an obvious constraint, but minimum funding levels or mandated capabilities [regardless of risk reduction] would not be obvious to the OR analysts).
4. A list of model capabilities, potential uses, and limitations.

8.17 Phase Two Modeling for Further Research

In further research under phase two, the model will be expanded to address all 21 capabilities, and a more comprehensive set of constraint equations will be developed to ensure that:

1. The appropriate "balance" is provided for funding different activities within a capability (e.g. ensure that RDT&E is adequately funded prior to purchasing equipment using OPN funding).
2. Inter-capability dependencies are established and adequate funding is provided as needed. The specific tasks for this phase include:
3. Develop a set of constraint equations used by the model, along with a brief statement that provides the basis for each constraint equation. A typical explanation might be a statement such as: "This constraint prevents purchase of chemical/biological detection equipment unless RDT&E activities are adequately funded". The constraints would be subsequently reviewed, potentially modified, and approved by CNI.
4. Provide the capability to use piecewise linear utility functions. The model will be upgraded to accept piecewise linear utility functions. These functions will initially be selected by the OR analysts. CNI will review, possibly modify, and approve the functions. The parameters used in the utility functions will need to be sufficiently easy to modify so they can be adjusted with minimal effort once they are approved by CNI.
5. Draft an explanation regarding how to interpret model results, including data that the model provides on sensitivity and marginal benefits associated with changes in funding level for a particular capability.
6. Provide the capability to retrofit the model with a time dimension (i.e. the capability to break the funding out over a six-year period with corresponding funding constraints).
7. Provide the capability to set funding constraints by total funding, type of funding, or both. CNI will have options on how to constrain the model. The model should have the capability to easily accept constraint equations by total funding for each of four types of funding (i.e. the type of constraint set used in the original model), total funding by year (in which case the time dimension capability must be enabled), or total funding over the six-year period (e.g. optimize resource allocation based on a total spend value over six years).
8. Expand the documentation of the model. In addition to providing an explanation of model results, document.
   - The constraint set, including the basis/function for each constraint.
   - Any assumptions made or data developed to test/exercise the model.
   - The purpose of each equation (or group of similar equations) in LINGO (i.e. provide comments to help facilitate future changes to the model)

8.18 Potential Model Expansion Preferences and Limitations

- Given the large number of constraints and rules that are anticipated, a full model may not have a valid solution. To the extent possible, the model should include constraint equations formulated in a way to help prevent this situation (e.g., expressed as inequalities or ranges around a particular input value).
- CNI will eventually set a constraint for each of the four types of funding for each FY between FY06 and FY11 (six years by four types of funding, resulting in 24 constraint equations involving funding), but it is likely that this effort will be iterative based on "what if" type of iterative analyses.
- CNI will probably want to optimize funding types (i.e. determine if changing from one type of funding to another within a FY will impact total risk reduction), but it is likely that this will be done based on a review of model results, particularly marginal benefits and sensitivity analysis data (e.g. the OR analyst will make manual adjustments and rerun the model rather than asking the model to determine an optimum funding strategy).
- CNI will probably want to perform sensitivity analyses on model constraints such as overall funding level, changing types of money, and moving funding from one year to another year

(presumably, they will only choose to model changes in funding from one year to the next if the time dimension is enabled).

- CNI will probably want to perform some sort of sensitivity analysis on the risk-related data that are used in the objective function, possibly via simulation.
- It is desirable to implement the model in some sort of interactive online tool that will facilitate iterative input-process-outputs exercises with different budget scenarios.

8.19 Solution 1

Solution obtained by setting a lower bound on RDT&E for security and surveillance of 0.25 and lower bound on OMN for security and surveillance of 0.1. Local optimal solution found at iteration: 6.

Objective value: 181.4502.

Variable	Value	Reduced cost
Y1	1.000000	733.4683
Y2	1.000000	0.000000
Y4	1.000000	0.000000
Y5	1.000000	0.000000
Y6	1.000000	0.000000
X12	1.000000	0.000000
X13	0.3000000	0.000000
X14	0.4000000	0.000000
X21	1.000000	0.000000
X23	0.5000000	0.000000
X24	0.2500000	0.000000
Y3	1.000000	0.000000
X31	0.2127660	0.000000
X33	0.1000000E–02	0.000000
X34	0.7420000	0.000000
X41	1.000000	0.000000
X43	0.6518429	0.000000
X44	0.2000000	0.000000
X52	0.9166667	0.000000
X53	1.000000	0.000000
X54	0.5000000	0.000000
X63	0.2500000	0.000000
X64	0.3000000	0.000000
X42	0.2500000	0.000000

Output of solution
Open all capabilities:

	MCON	RDT&E	OPN	OMN
Decision making	–	100%	30%	40%
Area wide alert 100%	100%	–	50%	25%
Operations center	21.3%	–	–	74.2%
Security surveillance	100%	25%	65.2%	20%
CBRN response	–	91.7%	100%	50%
C4I	–	–	25%	30%

8.20 Solution 2

Solution obtained by setting a lower bound on RDT&E for security and surveillance of 0.25 and lower bound on OMN for security and surveillance of 0.1, and relaxing requirement that contingency requirement on capabilities one and five. Local optimal solution found at iteration: 5.

Objective value: −14.584

Variable	Value	Reduced cost
Y1	1.000000	−16.10447
Y2	1.000000	0.000000
Y4	1.000000	0.000000
Y5	0.000000	0.000000
Y6	1.000000	0.000000
X12	0.1000000E−02	0.000000
X13	0.3000000	0.000000
X14	0.4000000	0.000000
X21	1.000000	0.000000
X23	0.5000000	0.000000
X24	0.2500000	0.000000
Y3	1.000000	0.000000
X31	0.2127660	0.000000
X33	0.1000000E−02	0.000000
X34	0.1000000E−02	0.000000
X41	1.000000	0.000000
X43	0.7028394	0.000000
X44	0.2000000	0.000000
X52	0.000000	0.000000
X53	0.000000	14.27902
X54	0.000000	0.000000
X63	0.2500000	0.000000
X64	0.3000000	0.000000
X42	0.2500000	0.000000

Open all capabilities except five:

	MCON	RDT&E	OPN	OMN
Decision making	–	–	30%	40%
Area wide alert 100%	100%	–	50%	25%
Operations center	21.3%	–	–	–
Security surveillance	100%	25%	70.3%	20%
CBRN response	–	–	–	
C4	–	–	–	30%

8.21 Conclusions

The modeling approach presented here is to illustrate how OR can be applied to complex budget allocation problems. The model can be adapted and modified for other budget decision applications. Other services (Air Force, Army, Marine, and Guard) can customize the basic modeling approach to fit their respective budget allocation practices and conventions.

References

Alberts, D.S., and R.E. Hayes. 2007. *Planning: Complex Endeavors*. Command and Control Research Program (CCRP) Publications Series, Washington, DC.

Alberts, D.S., P.M. Eggenhofer, P. Essens, R.E. Hayes, R.K. Huber, M. Moffat, M.E. Nissen, J. Romer, S. Schafer, M. Spaans, K. Titze, and W. Itreurniet. 2007. The future of C2. *The International C2 Journal* (special issue), 1 (1).

Badiru, A.B., and C.H. Aikens. 1996. LP Model of Budget Allocation and Optimization for U.S. Navy Installations, Operations Research Consulting Project Report, AB International Consulting Services, Knoxville, TN.

Badiru, A.B., and O.A. Omitaomu. 2007. *Computational Economic Analysis for Engineering and Industry*. Taylor & Francis CRC Press, Boca Raton, FL.

Damodaran, A. 1996. *Investment Valuation: Tools and Techniques for Determining the Value of any Asset*. John Wiley & Sons, NY.

Jones, C.P. 2006. *Investments: Analysis and Management*, 7 edn. John Wiley & Sons, NY.

Moffat, J. 2006. *Complexity Theory and Network Centric Warfare*. Command and Control Research Program (CCRP) Publications Series, Washington, DC.

9

An Overview of Meta-Heuristics and Their Use in Military Modeling

Raymond R. Hill
Air Force Institute of Technology

Edward A. Pohl
University of Arkansas

9.1 Introduction

The military employs a large number of models. The models vary tremendously by type and purpose. The more common types of models include: Guéret et al. (2003):

- Simulation models
- Optimization or mathematical programming models
- Financial models
- Logistics and inventory models
- Statistical models

An abbreviated list of purposes involving the use of these models include:

- Analysis of combat operations
- Analyses of weaponry effectiveness

- Development and analysis of deployment plans
- Programming for and analysis of weapons procurement
- Programming for and planning of deployment operations
- Personnel scheduling
- Logistics forecasting

Our present focus is on optimization models and the use of optimization modeling for a variety of planning and programming applications. An optimization model seeks some best allocation of resources, while meeting some set of resource limitations. Generally, a "best" solution is either minimizing or maximizing some cost or profit function, respectively. Constraints are defined and added to a model to capture any resource limitations. The process of optimization involves use of an algorithm that searches among potential allocations to locate a best feasible allocation. This search occurs using some computational representation of the optimization problem of interest.

Classic optimization algorithms use objective function derivative information to conduct an effective search among feasible allocations (i.e. searches within the feasible region of the problem defined by the problem constraints). These algorithms include the simplex algorithm for linear programs, the branch-and-bound algorithm for discrete optimization problems, and nonlinear algorithms such as the generalized reduced gradient for problems involving nonlinear functions. These approaches provide guarantees of convergence to global optimal solutions for linear models or local optimal solutions for nonlinear models. We will refer to these as mathematical programming approaches.

Our precise focus in this chapter is on the use of heuristic search methods applied to optimization problems with a military focus. We can distinguish heuristic search techniques from mathematical programming techniques by the following characteristics. A heuristic algorithm:

- Does not guarantee convergence to a global optimal (in general);
- Does not employ gradient-based techniques;
- Involves some form of local neighborhood evaluation; and
- Has some analogy to a heuristic problem solving approach or an analogy to some natural process

The remainder of this chapter is organized as follows. In the heuristic search concepts section we provide a fundamental understanding of important concepts pertaining to heuristic search methods for optimization. These concepts provide a foundation for subsequent sections that overview various heuristic search techniques. Subsequent sections provide overviews of simulated annealing (SA), genetic algorithms (GAs), tabu search, scatter search and ant colony optimization (ACO). In each of these sections we introduce the motivations for the techniques, the fundamental concepts pertaining to how the technique functions and when appropriate discuss some of the advanced concepts associated with the technique. In the military applications section we provide an overview of how some of these heuristic search methods have been applied to military-focused modeling applications. We close with concluding remarks pertaining to the use heuristic search methods for military optimization applications.

9.2 Heuristic Search Concepts

As Glover (2007) notes, "[a] popular thrust of many research initiatives, and especially of publications designed to catch the public eye, is to associate various methods with processes found in nature." This has been the allure and sometimes the bane of heuristic and meta-heuristic search methods for optimization. The more recent contributions to heuristic search optimization seem to have a focus on analogies to naturally occurring phenomena. As Glover notes, there "exudes a certain mystery" regarding how this real-world phenomena embodied within some computational algorithm can solve complex optimization problems, often with results as remarkable as those attained natural phenomena upon which the algorithms are based. What is sometimes lost however is that the real value of these analogous methods is to understand the motivation for the naturally occurring problem solving capabilities, the context within which the behavior is found effective, and then to comprehend how to instantiate such behaviors in computational algorithms applied in the context of solving optimization problems. Thus, these natural metaphors are useful so long as the metaphor does not prevent productive scientific inquiry (Glover 2007).

From Silver (2004), "the term heuristic means a method which, on the basis of experience or judgment, seems likely to yield a reasonable solution to a problem, but which cannot be guaranteed to produce the mathematically optimal solution." The use of heuristic methods is not new to optimization applications. Early research examined a variety of heuristic solution methods for solving hard discrete optimization problems. Most optimization courses consider heuristic methods such as the Hungarian method, the nearest-neighbor route construction approach, the minimum spanning tree algorithm, or the cost-benefit ratio calculations for knapsack problems. These heuristics were based on selecting decision variable values that produced some sort of best immediate gain based on the current information available to the algorithm, thus a common label "greedy" was applied to these approaches. These approaches were greedy in the sense that current decisions were based just on currently available data versus basing decisions on globally available information.

The problem with greedy heuristic approaches is sometimes a poor quality solution results. A pitfall of the greedy approach is that "good" variable selection and valuation earlier in the solution construction process is usually offset with "poor" variable valuation later in the process due to the inescapable limitations on the variable selections; the greedy approaches suffer from a loss of freedom of choice later in the process. This pitfall is often referred to as getting "trapped in regions of local optimality" and the regions where the heuristic search is trapped too often contain poor solutions.

The focus of this chapter is on the "modern" heuristics which contain mechanisms to first construct a solution and then improve upon that solution by including mechanisms to escape from the locally optimal regions. These modern heuristics, many of which were discovered when examining algorithms based on some natural phenomena, are collectively referred to as meta-heuristics. Borrowing from the definition of Osman and Kelly (1996), a "meta-heuristic is an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting the search spaces using learning strategies to structure information in order to find efficiently near-optimal solutions." The nature of how the subordinate search is guided, how the solutions are structured and how the process learns varies by heuristic and specifics of each are covered in their individual sections in this chapter. The remainder of this section provides some of the common terminology and concepts underlying meta-heuristic search strategies.

Any meta-heuristic seeks good solutions to the optimization problem. These solutions contain those combinations of decision variable values that provide the best available solution for some specified evaluation criteria while meeting all restrictions on the solution. All solutions have attributes or characteristics that define the solution. For instance in routing problems, typical attributes are the number of routes, number of vehicles, the customers assigned to each route and the order the customers are visited within each route. In assignment problems, attributes might be how jobs are assigned to workstations, the order of the jobs within the workstations, when jobs need to be completed, how long jobs require for completion, and when jobs should complete. The solution evaluation is that function that maps the combination of decision variable values to some numerical indicator of quality; it is the objective function for the problem. The solution restrictions are the constraints on the problem. A meta-heuristic systematically searches among problem solutions moving among those solutions via some transition function. A transition function is defined for the problem and provides the basis for changing a current solution to a new solution thus constituting a move among the solutions in the problem solution space. Finally, meta-heuristics exploit the neighborhood concept to control the search. The neighborhood of a solution consists of those solutions that can be obtained by an application of the transition function. An improving search continues to find better solutions through the search process. As the rate of solution improvement drops, the search is said to have slowed. When search improvement stops, the search is said to have stagnated, stabilized or in some cases to have converged.

More formally, let Ω be the set of solutions to the optimization problem of interest, P. Define $\Omega^f \subset \Omega$ to be the set of feasible solutions to the optimization problem and $\Omega^{\text{inf}} = \Omega \backslash \Omega^f$ be infeasible solutions. Then $x \in \Omega$ represents a solution to the problem P and $f(x) : \Omega \mapsto \Re$ represents the evaluation function that maps that combination of decision variables to some real-valued indicator of the quality of that solution. The transition function, $t(x) : x \in \Omega \mapsto x' \in \Omega$. The neighborhood function is thus defined as $N(x,t(x))$ and can be thought of as a subset of Ω "in the vicinity" of the current solution x (Osman and Kelly 1996).The neighborhood of a solution x will be denoted simply as $N(x)$. The mathematical formulation of the problem P is thus

$$\text{Optimize } f(x)$$
$$\text{Subject to } x \in \Omega^f.$$

Note that while the goal of a heuristic search algorithm is to return some best feasible solution, there is no restriction that the search itself remain absolutely feasible throughout. This is a distinguishing characteristic of heuristic search methods and in many cases provides the heuristic tremendous efficiency in eventually finding good feasible solutions.

The general approach for a heuristic search algorithm is the following:

1. Initialize any search parameters; generate an initial solution
2. Generate $N(x)$
3. Evaluate members of $N(x)$
4. Select some $x' \in N(x)$
5. Set $x \leftarrow t(x')$
6. If not done searching, Go To Step 2, else return x as final solution

This general approach is extended in actual practice. For instance, a variety of procedures might be used to generate an initial solution. In some cases, the algorithm may generate a candidate list of potential solutions. This candidate list of solutions is accessed whenever the search process is deemed to have slowed, the next member of the candidate list is set as the new initial solution, and the search is re-started. For some algorithms, the definition of the neighborhood may change dynamically during the algorithm or the neighborhood itself may be restricted to avoid examining large neighborhoods. Muller-Merbach (1981) provides a thorough examination of heuristic search neighborhoods. Evaluation of the members of a neighborhood can be a computational expensive process so some algorithms employ strategies to reduce the neighborhood evaluation effort. Examples of neighborhood strategies include:

- Evaluating all members and choose the best
- Evaluating members until the first feasible solution found
- Evaluating members until a first improving solution is found
- Evaluating a subset of members choosing the best in the subset; or
- Evaluating and use some randomly chosen member

Finally, practical heuristic implementations employ data structures as a means to store a set of best-solutions found during the search. This set of best solutions found can then be used for such things as restarting the search process, providing a range of potentially good solutions, or providing that best solution within the set representing the best solution found by the heuristic search algorithm.

A key aspect of heuristic search algorithms is their explicit strategy to escape regions of local optimality. Classic methods, such as linear programming, gradient descent (or ascent) for nonlinear problems, even greedy heuristic search algorithms stop when there are no improving moves. In the case of linear programming, the solution returned is optimal. In the case of nonlinear problems, the solution returned is a local optimal. For greedy heuristic search algorithms, lack of an improving move stops the process and the process is deemed to be at a local optimal (albeit often not a very good local optimal). The power of modern heuristic search algorithms is their ability to escape these regions of local optima and continue transiting through the search space. Some of the more common escape mechanisms include:

- Acceptance of nonimproving moves as a means to transition the search into new regions of Ω.
- Re-starting of the search from a pre-defined set of solutions (i.e. using the candidate list) or from a newly generated solution.
- Generation of a random solution and a transition from the current solution to this new solution.
- Modification of the transition function such as in variable neighborhood search strategies (Hansen and Mladenović 2003).

We close this background section with definitions of other terms commonly found in the heuristic search literature. First, diversification within a search algorithm refers to the systematic exploration of Ω by means of moving into new, potentially unexplored regions of the search space. The strategies employed for diversification vary by algorithm considered. Conversely, intensification within a search algorithm refers to a re-inforcement of the local improvement strategies to move toward locally optimal solutions. The more strategic search algorithms will incorporate mechanisms transitioning between periods of intensification and periods of diversification.

The concept of learning is often used in the discussion of a heuristic search algorithm. In general, learning entails using mechanisms within the search that tune the search based on the results of the search. In some algorithms, for instance SA and GAs, these mechanisms might involve changes to the search parameters. In other algorithms, like tabu search or scatter search, these mechanisms might involve the strategic use of data structures to store information regarding search progress and then using that data, the memory of the search, to modify specific aspects of the search. Finally, a dynamic search algorithm is one that can change or adapt during the search to make the search more productive and efficient.

9.3 Simulated Annealing

9.3.1 Background and Motivation

SA is one of the earliest search heuristics for optimization inspired by a natural analogy. The analogy is "to the process of physical annealing with solids, in which a material is heated and then allowed to cool very slowly until it achieves its most regular possible crystalline state, with corresponding minimum energy" (Henderson et al. 2003).

The typical example discussed for this annealing analogy concerns the forming of crystals. To avoid causing defects in the final crystal structure, the substance is melted and then cooled slowly. During the cooling process a long time is spent at each temperature level particularly those temperatures closer to the freezing point. This cooling schedule allows the substance to reach an equilibrium state at each temperature so that the final ground state is void of defects.

As discussed in Kirkpatrick et al. (1983), Metropolis et al. first introduced a computational form of SA as a means to simulate "a collection of atoms at equilibrium at a given temperature." The Metropolis approach involved the probabilistic acceptance of nonimproving steps of a process. In their case, when an atom is displaced causing a decrease in system energy, the displacement is accepted and the simulation proceeds. If the energy is increased then the displacement can be rejected with some probability. That probability function involves a parameter called the temperature. This temperature parameter decreases during the execution of the algorithm causing changes to the probability of accepting the nonimproving moves.

Kirkpatrick et al. (1983) tied the Metropolis algorithm to optimization and demonstrated the algorithm on various optimization problems.

9.3.2 The Algorithm

SA is a local search algorithm with a probabilistic element. It is this probabilistic element that provides the SA with the escape mechanism to depart the local region and diversify the search early in the algorithm process, a mechanism that diminishes over time to provide intensification of the search later in the algorithm process.

The SA search process, assuming we want to find a maximum value, is:

1. Initialize the search process
   - Set initial solution x_0
   - Set temperature reduction factor, $0<\alpha<1$, initial temperature, f_0, and f_{min} as minimum temperature.
   - Set *nreps* as maximum iterations per temperature and *reps* = 0 as current iteration counter.
   - Set $x = x_0$ and $f = f_0$.
2. Randomly select $x^0 \in N(x,t)$. Increment *reps*.
3. Calculate $\delta = f(x) - f(x^0)$.
4. If $\delta > 0$ then set $x = x^0$ else generate $u \in U(0,1)$ and then if $u < \exp(-\delta/f)$, $x = x^0$.
5. If $reps \le nreps$, go to Step 2.
6. Set $f = \alpha f$, *reps* = 0. If $f \ge f_{min}$, go to Step 2, else Return x as final solution.

In Step 2, the transition function randomly selects a neighbor of the current solution x. The change in solution is calculated in Step 3. In Step 4, improving solutions are always accepted while nonimproving solutions are selected with probability that is a function of the current temperature, f. When the temperature is high, the probability of selection is higher mimicking the molten state of the physical annealing process. As the temperature decreases the probability of selection decreases to mimic the cooling process. At high temperatures the SA algorithm resembles a random search while at low temperatures the SA algorithm resembles a descent function. The *nreps* parameter provides a number of iterations at a temperature level mimicking the time the physical annealing process spends at a temperature (Step 5). In Step 6, the temperature is lowered by the factor α. When α is high, the process mimics slow cooling while α low mimics rapid cooling. When the temperature reaches some pre-defined minimum, f_{min}, the SA process stops and returns the current solution, x.

As with any heuristic the SA algorithm requires problem specific decisions prior to implementation and execution. These decisions are:

1. A formulation of the problem.
2. A neighborhood definition and transition function.
3. An objective function.
4. A strategy for handling any problem constraints.
5. Some procedure for obtaining an initial solution.

The SA can actually be thought of as a set of algorithms each distinguished by the parameter values assigned. These more generic algorithm specifying decisions include setting the values for α, f_0, f_{min}, and *nreps*. Unfortunately, few guidelines exist for setting these parameters to their best values; generally, the parameter values are empirically tuned on sample problems and then set to the derived best values for solving actual problems.

9.3.3 Algorithm Enhancements

The fundamental tradeoff in an SA is the cooling rate, α, and the number of iterations at each temperature, *nreps*. These together form the cooling schedule for the algorithm. Theory suggests using large values of α and large values of *nreps*. This combination however can lead to computationally expensive procedures. Practical implementations choose smaller values, in some cases even setting $nreps = 1$ offset by large α values.

While the strict SA analogy saves the current solution only, optimization enhancements include providing data structures that save the best solutions encountered during the search. This list is surveyed when the algorithm execution completes and the best solution in this list is returned. The SA analogy also precludes increasing f but the enhancement called reheating involves the re-setting of f to a higher temperature to promote additional diversification in the search process.

Additional enhancements are more problem specific. For instance, transition functions and neighborhoods can be adjusted as a function of temperature to induce diversification or intensification. Restricted neighborhoods can be used to accommodate problem constraints or to improve processing time. Data structures can be exploited to save the search history and thus avoid re-selecting previously selected solutions. Finally, the SA can be hybridized with other search methods such as steepest descent (or ascent) to further improve solutions prior to returning the final solution from the algorithm.

9.3.4 Final Comments

SA is one of the earliest meta-heuristics. As such, it has been used in many applications and has achieved success in complex optimization settings. Recent research has demonstrated that SA may not be a preferred approach for many classes of problems; other heuristics tend to outperform the SA on these problem classes. However, SA was one of the first meta-heuristics to have convergence results thereby helping to legitimize meta-heuristic use in optimization applications. While often viewed as a random search, SA is really a neighborhood search that interleaves varying degrees of random escape mechanisms with local improvement to achieve reasonably good solutions to hard (discrete) optimization problems.

9.4 Genetic Algorithms

9.4.1 Background and Motivation

Genetic algorithms (GAs) are search procedures inspired by biology and the workings of natural selection. Conceived by John Holland and his colleagues, GAs have been applied in many diverse applications (Chu and Beasley 1998). The GA name "originates from the analogy between the representation of a complex structure by means of a vector of components, and the idea, familiar to biologists, of the genetic structure of a chromosome" (Reeves 1993). In biology, natural selection reinforces characteristics most amenable to a species survival. Genes within the chromosomes of the stronger members, corresponding to the more desirable characteristics, pass to subsequent generations through the reproduction process.

When applied to optimization problems, consider solutions to problems to be members of a population. Associate with each member of this population some value that equates to the strength of the solution, for instance the objective function value. We then provide for a reproduction process among members of the population to occur producing new members, or solutions, to the problem but bias this reproduction process so that the ability of a population member to reproduce is a function of that member's value. This means better solutions, or stronger population members, provide greater influence on future solutions, population members for the next generation, than do weaker solutions. To make this work we encode the problem solutions such that problem characteristics can be passed onto subsequent population members. Starting with a population of solutions and allowing a "survival of the fittest" evolution process to occur over some number of generations will produce a final population of strong solutions and ideally the optimal solution from within that population.

9.4.2 The Algorithm

A GA commences with an encoding of problem solutions. A typical encoding scheme is to represent the solution as some binary stream. This binary stream is referred to as the chromosome. Sections of this chromosome representing logical components of the solution are called the genes; for instance, each decision variable is represented by some binary stream (the gene) which collectively concatenated yields the chromosome. The actual representation of the problem solution is referred to as the phenotype while the encoded representation is referred to as the genotype. The mapping from phenotype to genotype is often problem specific and user defined.

Each solution must have some value or fitness function defined for it. This fitness function, evaluated in the phenotype is associated with the genotype and exploited during subsequent evolutionary processes within the GA. This fitness function must account for objective function values and solution feasibility; quite often penalty functions are used to de-value infeasible solutions as a function of their distance from feasibility.

The GA is a population based heuristic. The GA starts with some initial population and evolves that population through subsequent generations. One way of generating an initial GA population is to generate and evaluate random solutions and place those solutions into the initial population. Other methods involve generating a variety of initial solutions via constructive heuristic approaches. Hill and Hiremath (2005) use an intelligent approach to randomly generate initial population generation members of high solution quality to improve overall GA performance.

A GA functions based on analogies to evolution, which involve population dynamics. Each specific population instance is called a generation. The evolutionary process allows each generation to produce off-spring that potentially replace individuals in the current population. This is called the GA birth and death process. In each generation, genetic operators are applied to selected individuals from the current population to create off spring that lead to a new population. The result of these population dynamics is a changing of the makeup of the population with each generation; each subsequent generation should be stronger than the previous generation. In optimization terms, each new generation should have a higher average value associated with the population members and the "best" member of the population should never be worse than the best member of the previous generations. These are key attributes for a GA. When the average solution value of the generations or there is no change in the fitness of the best found solution, the GA is said to have converged. Intuitively it means the population is unable to evolve newer, stronger members, and thus the population is unable to evolve any better solution that what has already been found.

Figure 9.1 (Renner and Ekart 2003) shows a general flowchart of a GA.

Three main genetic operators are used in the population dynamics: reproduction, crossover and mutation. Varying probabilities of applying these operators can control the speed of convergence of the GA. The choice of crossover and mutation operators are most important to the performance of the GA. Hence, crossover and mutation operators must be carefully designed.

Reproduction is a process in which a new population is created by copying individual strings from the present population according to their fitness function values (Goldberg 1989). Copying strings according to their fitness values means that strings with higher fitness values have a higher probability of contributing one or more offspring for the next generation. This gives the possibility of survival for already developed fit solutions, an element of immortality to the problem solution or population member. There are a variety of ways to select population members for reproduction:

- Random: Pick two elements of the population randomly and allow these two to serve as parents.
- Rank: Rank order population members and pair parents by ranking.
- Elitist: Only choose among the strongest population members to serve as parents.
- Tournament: Pick a random subset from the population, put the best of the subset into the pool of parents allowing multiple copies of any population member.
- Proportional: Normalize population member fitness functions and randomly pick parents using the normalized fitness as their probability of selection; allow multiple copies of any population member.

New individuals are created as offspring of two parents via the crossover operation. One or more crossover points are selected at random within the chromosome of each parent. The parts that are delimited by the crossover points are then interchanged between the parents. The individuals formed as a result of

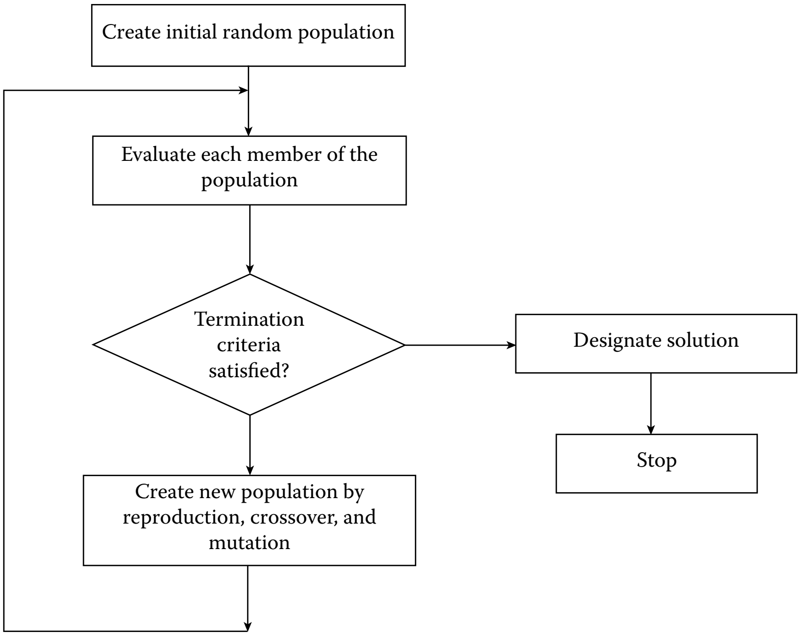


FIGURE 9.1 Genetic algorithm flowchart. (From Renner, G. and A. Ekart. 2003. Genetic algorithms in computeraided design. *Computer-Aided Design* 35, 709–726. With permission.)

such an interchange are the offspring (Zalzala and Fleming 1997). Burst crossover is a type of crossover where the crossover is made at every bit position, thus possibly resulting in many crossover-points, depending on the selected probability (Hoff et al. 1996). Common types of crossover are depicted in Figure 9.2 (Zalzala and Fleming 1997; Renner and Ekart 2003).

It is more desirable to keep fit individuals intact in the later phases of evolution. Hence it is a good idea to use an adaptively changing crossover rate (higher rates in the early phases and a lower rate at the end of a GA). In some cases it is also desirable to use several different types of crossover at different stages of evolution as shown in Figure 9.2 (Zalzala and Fleming 1997; Renner and Ekart 2003).

The final reproductive operator is the mutation operator. Mutations are random changes made to the off-spring chromosome. For instance, for a binary encoded chromosome, each bit in the string is examined and with some small probability, its value is complemented. Mutation provides a element of diversification in the GA plus serves the important task of restoring potentially good solution characteristics that might otherwise get lost during the crossover operations.

There are a variety of choices in terms of how generations evolve. The production of offspring expands the size of the population, or the current generation. A GA retains a constant population size, thus some of this expanded population must be removed, or killed off, to restore the population to its proper size. Once restored, the resulting population represents the next generation of the GA. Some of the ways populations are restored are:

- Generational: Offspring replace parents yielding a new population.
- Incremental: Proportionally select members for the next generation from the current, expanded population.
- Best: Only keep the strongest members of the expanded population.

The GA continues for some defined number of generations, or until convergence of the population is concluded. At that point, the best solution in the population is returned as the estimated "optimal" solution to the problem.

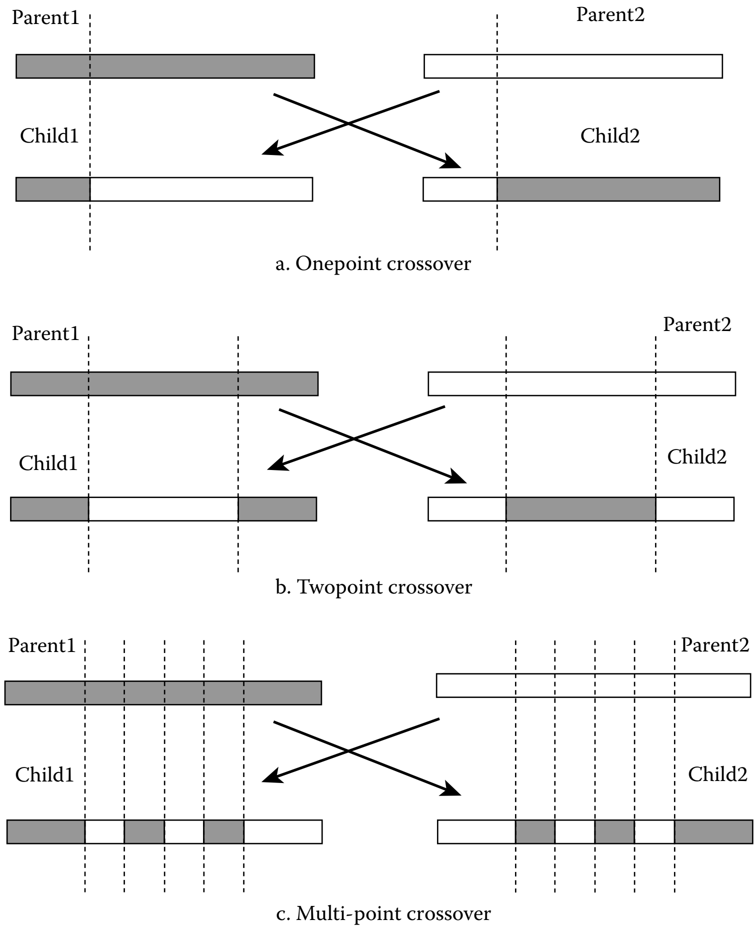


FIGURE 9.2 Different types of crossover operators. (From Zalzala, M.S. and P.J. Fleming (Eds.). 1997. *Genetic Algorithms in Engineering Systems*. London: Institution of Electrical Engineers. With permission.)

9.4.3 Algorithm Enhancements

There have been a variety of enhancements to the basic GA described above. Only some of those enhancements are discussed here. More complete information can be found in the various texts available on GAs. The interested reader is also directed to Illinois Genetic Algorithms Laboratory (http://www.illigal.uiuc.edu/web/) for some of the latest developments in GAs.

As with most heuristics based on some natural analogy, the underlying analogy is relaxed to implement processes conducive to improved optimization performance. The first such improvement was to ensure the best member, or members, of a population survive into the next generation. Dynamically changing problem structure can help the search process as well. For instance, increasing the mutation rate can be used to diversify a stagnant search. Re-starting the population provides a means to re-start the search. Expanding the population size for some period of time is another way to increase diversity. Changes to population dynamics operators include using complicated crossover patterns, fixing certain sections of chromosomes to not allow destruction during crossover, and ensuring that off-spring are feasible solutions by using repair operators that ensure the phenotype is feasible.

The reproduction process can cause irregularities in subsequent populations. For instance, a population can have perfect twins; two solutions are exactly the same. One adjustment to a GA is to ensure all population members are unique. A population member may not be a particular strong member, but is easily transformed into a stronger member. A Lamarchian process takes solutions from the population, applies a local search heuristic to that solution, and if the solution is improved, replaces that solution in the population with the improved solution.

There are also ways to improve the initial population. Elements of a randomly selected population can be improved via local improvement heuristics. A randomly selected population can be improved by intelligently modifying the probabilities associated with the solution selection process (Hill and Hiremath 2005). The initial population can also be seeded with strong methods found via other optimization methods or via specific domain knowledge about the problem. Typically, the stronger the initial population of a GA the quicker that GA can converge and return its recommended solution.

9.4.4 Final Comments

The GA is a popular meta-heuristic solution approach. Because the GA works on an encoded version of the optimization problem, the algorithm is applicable to a wide variety of problems; once encoded with a suitable fitness function available, the GA is quite problem independent. As a result the GA is arguably the most general purpose meta-heuristic and also arguably one of the two most popular.

The GA does have distracting elements. Some, such as Ross (1997) view the GA as an inferior approach. Since the GA works on a population of solutions, each of which require evaluation, the computational overhead can be troublesome. In some applications, nonbinary encoding schemes are required and crossover operations require a good deal of problem-specific knowledge. In such cases the GA label may be inappropriate. The GA can also experience convergence problems when later generations become "too similar" to effectively produce strong, new solutions. As with most meta-heuristics, the GA employs algorithm parameters that may need tuning to work effectively on the problem at hand. This requires additional investigation before one can actually solve the problem of interest. Finally, understanding the schema theory (Reeves 1993) proposed to explain GA performance is not suited for casual reading.

Additional details can be found in Beasley et al. (1993a, 1993b), the article by Reeves (1997) and the overview chapter by (1993).

9.5 Tabu Search

9.5.1 Background and Motivation

Tabu search's analogy is to mimic the human problem solving process. As Glover (1998) recounts one of his graduate school projects involved examining how people would solve an IP. Glover observed how the human problem solving process involved memory to help avoid repeating self-defeating sequences and involved radical changes to their approach when the current search strategy stagnated. Fast forward to the 1980s and the tabu search framework with problem specific search mechanisms make use of data structures (memory) to both intensify and diversify the search process. As Glover (1986) notes:

> The emergence of strategies that can be adapted to take advantage of diverse problem conditions marks what I believe to be the wave of the future of IP, giving us power to solve a significantly increased number of IP problems effectively.

9.5.2 The Algorithm

The bane of local search heuristics is the basin of local optimality within the often highly irregular search landscape; once the search process enters the basin, the local search heuristic can remain trapped in that basin. The tabu search framework introduced the concept of the tabu list, a data structure of solution attributes deemed "off limits" for some number of search moves (called their tabu tenure). Ignoring tabu attributes causes good features to remain in solutions and provides a mechanism to introduce new features thus allowing the search to move out of regions of local optimality. Use of multiple tabu lists provides further richness to the search process, at the expense of some additional processing.

The aspiration criteria allows the search process to override tabu restrictions and visit strong solutions examined within the current neighborhood that might otherwise be ignored. As with tabu lists, multiple aspiration criteria can be employed during a tabu search process.

In a 1977 paper, Glover (1977), inspired by Senju and Toyoda (1968), introduced the concept of strategic oscillation. Strategic oscillation allows a search process to systematically cross barriers other search approaches strictly apply. The most common barrier for IP problems is feasibility, either feasibility due to the constraints or feasibility due to the integrality restrictions. A systematic path through the infeasible region coupled with a search path back into the feasible region can compel the tabu search to reach new, potentially favorable regions of the search space.

The basic tabu search framework can be summarized as:

1. Initialize.
   - Determine search neighborhood(s) to be used and solution evaluation criteria.
   - Define tabu list(s) structure(s) and processes.
   - Define aspiration criteria and memory structures to employ.
   - Create initial solution to problem.
2. Local search. Examine and evaluate neighboring solutions. Select best nontabu neighbor or best tabu neighbor that meets aspiration criteria.
3. Implement move to the selected neighbor. Update tabu list structures (and other memory structures used). Return to Step 2 if stopping conditions are not met.
4. Return best solution found.

The basic tabu structure outlines above can be expanded to include pre-defined strategic oscillation phases. This structure can also involve dynamic aspects. Battiti and Tecchiolli (1994) dynamically change the tabu tenure employed based on recorded search progress. In their reactive tabu Search, if memory structures indicate the search progress has slowed or is cycling among sequences of solutions, the tabu tenure is (temporarily) adjusted to diversify the search.

9.5.3 Algorithm Enhancements

The basic tabu search outlined above provides a flexible, robust, and generally effective search process. However, on more difficult problems, additional structure can be exploited to try and improve the solution returned from the basic tabu search. Within the tabu search framework, these additional structures comprise intermediate and long-term memory approaches.

9.5.3.1 Intermediate-term Memory Approaches

The key component of intermediate-term memory is the elite list. This list contains good solutions encountered during the search process. For example, the list may contain good neighbors evaluated but not selected for a move during the basic tabu search process. Kinney et al. (2005) build an elite list via a variety of solution construction methods for routing problems.

When invoked during the tabu search process, intermediate-term memory search efforts will extract a solution from the elite list and re-start the tabu search from that solution. This re-start process may or may not alter the current status of any memory structures to tabu lists; this is a strategic choice made by the tabu search designer. The extraction and re-start process continues using elements from the elite list for a pre-determined period of time or until the elite list is exhausted.

Placing solutions on the elite list is tabu search specific; the search design process determines the membership criteria. Examples of criteria include: solution is some percentage from overall best solution, some percentage from best solution in the current neighborhood, or is among some set of top solutions in the neighborhood. Use of the elite list diversifies the search into new regions and intensifies the search by using good solutions from the elite list to re-start the search process.

9.5.3.2 Long-term Memory Approaches

The long-term memory aspects of tabu search are intended to radically diversify the search and involves search processes that can be fairly complex. The long-term memory aspects of the tabu search framework are unique in the realm of heuristic search methods.

The vocabulary building process involves the structured use of memory to build new solutions from which the tabu search process can re-start. During the search process, as moves are implemented, a catalog of solution attributes can be maintained. This catalog can then be accessed at some point during the search process to ascertain attributes often used and attributes infrequently used in solutions encountered during the tabu search process. Such attributes can then be combined to create new solutions. For instance, using infrequently found attributes can yield a very new solution. The solution might be of poor quality but may be so diverse a solution that in places the tabu search process into a new region of the search space, a region that may have really good solutions contained within it. Combining frequently used attributes may yield a very strong solution. Sometimes the solution may be infeasible but may lead to a search trajectory back to the feasible region that ends up finding a new region to explore. The structure of the attribute catalog, how the catalog is maintained, and how the catalog is used to create the new solutions is problem specific and generally a product of the advanced learning mechanisms of tabu search.

Path relinking involves the use of good solutions and a constructed search along the path connecting the two solutions. This involves a problem-specific routine that transforms one good solution to another good solution via a sequence of attribute changes. One rationale for the path relinking approach is that better solutions may lie between good solutions and local search mechanisms may be unable to follow the path between the good solutions using just the defined neighborhood search functions.

Controlled randomization can come into play in the tabu search framework and this involves modifying the search neighborhood such that some set of good solutions, from the current neighborhood, are selected and then one of the elements from this set are picked probabilistically. For instance, the approach may be to take the top five neighbors of a current solution. To maintain the aggressive nature of the tabu search process, we then bias the random selection process by assigning each element of the set a probability of selection that is a function of their solution quality (i.e. objective function value). The stronger neighbors thus have a better chance of selection. The random component adds an extra element of diversity to the search process.

Each of the above long-term memory strategies are evoked based on rules embedded within the tabu search process. Such rules generally emanate from a target analysis effort performed on the problem type. Target analysis is the key learning process of tabu search. It involves a systematic off-line study of solution search performance, a conjecture regarding new search rules, and experimentation to support or reject the conjecture regarding the candidate rule. Effective search rules are subsequently embedded into the tabu search framework for use on general instances of the problem.

9.5.4 Final Comments

Tabu search features three key themes (Glover 1990):

1. Flexible memory structures.
2. A mechanism to control the freeing and constraining of the search process.
3. Memory structures of differing time spans to promote intensification and diversification.

The resulting framework of tabu search provides a flexible, powerful mechanism for solving IP problems of varying complexity. The flexibility of the framework has yielded a variety of innovative optimization search processes. Each of these search processes feature the interplay between intensification of the search in promising areas of the search space and then diversification of the search into new regions of the search space when the search process is deemed to have slowed or even to have stalled.

Details of the tabu search framework are plentiful. However, the interested reader can find the key insights in Glover (1989, 1990a, 1990b), the text by (Glover and Laguna 1997) and Glover's chapter in (Reeves 1993).

9.6 Scatter Search

9.6.1 Background and Motivation

Scatter search was first proposed by Fred Glover in a 1977 paper (Glover 1977) and was then effectively shelved until 1990 (Laguna and Martí 2003).The motivation for the concept came from formulations that obtained solutions by combining decision rules and constraints (see Laguna and Martí 2003 or Glover et al. 2000 for details).

Another way to consider scatter search is to assume that given some set of good solutions to a problem, potentially better solutions may exist in the subspaces spanned by subsets of these good solutions. In a simple example, given two good solutions, x^1 and x^2, consider potential solutions on the line containing both x^1 and x^2.

The 1977 paper by Glover (Glover 1977) should be read by anyone interested in heuristics because of its subsequent influence. Not only did the paper lay out the initial scatter search, the paper presented some of the key ideas later found in Glover's tabu search. Some key concepts introduced in tabu search are also found in scatter search.

9.6.2 The Algorithm

The basic scatter search has changed since its initial proposal. The current basic scatter search template has five methods (Laguna and Martí 2003).

The diversification method is defined as a way to generate a fairly large set of trial solutions. Typically some seed solution is generated and the diversification method is used to create the trial pool with sufficient diversity to adequately cover the solution space.

An improvement method is used to transform some trial solution into an enhanced (i.e. better) solution. This method will naturally vary by problem type.

A reference set update method is used to maintain some "best" set of solutions of high quality and sufficient diversity. This best set is called the reference set.

A subset generation method extracts members of the reference set to create new combined solutions for the problem of interest.

A solution combination method uses the results from the reference set to create the combined solutions.

These template methods are used to create the specific scatter search implementation. The generic scatter search process is defined as the following four step process:

1. Step 1. Generate starting set of solutions; improve solutions; extract reference set.
2. Step 2. Create new pool of solutions via solutions combination method.
3. Step 3. Attempt to improve the solutions in the new pool using the improvement method.
4. Step 4. Update reference set with best solutions from the improved pool. Return to Step 2 until stopping condition is met.

Upon completion the final reference set will contain the best solution returned by the algorithm. Additionally, a dedicated memory structure can be employed to specifically retain the best or set of best solutions encountered during the search process.

9.6.3 Algorithm Enhancements

While the basic scatter search is quite flexible, additional complexities have been added to improve he power of the algorithm particularly when used on difficult problems.

The basic scatter search employs a static method for reference set updating. This static approach examines all combinations from the reference set to create the trial pool from which the reference set is updated. A dynamic update method puts quality trial solutions into the reference set and makes those new reference set members available for the combination method. This quickly removes poor solutions from the reference set and gets the better solutions involved in the search process sooner.

When scatter search progress slows, the reference set has become ineffective. The re-building of the reference set can be triggered by a stabilization of the search progress. One way to rebuild the reference set is to create a new reference set with maximally diverse solutions (Martí et al. 2006).

The reference set can be tiered to avoid search stagnation. While the reference set is generally updated based on solution quality, a two-tier reference set would add updates based on solution diversity as well. Thus, the reference set will use high quality solutions and diverse solutions in the combination generation method. The intent is to keep the search aggressive (with the high quality solutions) and productive (with the diverse solutions).

Typically, combination methods consider reference set solutions in pairs. Advanced combination methods can employ three or more reference set members to create a new combination. This use of multiple members, when coupled with improved combination methods, provides tremendous diversification capability for the scatter search process.

Finally, memory structures can enhance the search. Hashing functions can be used to store all solutions encountered during the search thereby avoiding the re-examination of solutions. Attributes of solutions can be tracked using data structures and the information exploited to generate diverse solutions or to intensify the search. Such strategic uses of memory is influenced by the success of these approaches in tabu search.

9.6.4 Final Comments

Scatter search is still a relatively new meta-heuristic search technique. Martí et al. (2006) provide search engine results indicating scatter search still trails tabu search and GAs in terms of influence. However, scatter search is easily hybridized with other methods making the approach quite flexible. The scatter search technique is also quite adept for simulation optimization applications and is found in the OptQuest optimization package (Laguna and Martí 2003). Thus, the future of scatter search is rich with research avenues and applications.

9.7 Ant Colony Optimization

9.7.1 Background and Motivation

Ant colony algorithms are inspired by the foraging behavior of a colony of ants. Despite their lack of communication skills (as we know them), foraging ants demonstrate the ability to collectively arrive at good solutions to the routing problem they face finding a shortest path from their nest to a food source. In fact, they demonstrate an ability to adapt their routes when those routes get disrupted.

As Dorigo et al. (1996) discusses, the ants employ an indirect form of communication wherein ants deposit pheromone along their selected trail. Shorter trails are transited quicker, and thus more often and so are re-inforced with more pheromone prompting subsequent ants, that sense the pheromone deposits, to follow those quicker paths. This stigmergistic behavior is mimicked in optimization algorithms. For the optimization algorithm the path length is used to re-inforce the segments selected for the path; shorter paths receive more re-inforcement than do the longer paths. The challenge is to characterize the optimization problem in terms of segments over which the ants travel to complete a solution to the problem.

9.7.2 The Algorithm

The ant colony optimization (ACO) algorithm computationally mimics the environmental re-inforcement behavior observed in real ants. The ACO algorithm requires the following elements.

- Optimization problem mapped into the structure of a set of nodes, connected via a sequence of states, or segments, that an ant travels over to construct a solution to the problem.
- A means to evaluate the quality of the solution obtained by the ant.
- A data structure used to track the artificial pheromone deposited (and remaining) on each of the problem segments.
- A means for updating the pheromone data structure.
- A transition policy employed by the ants to choose among segments while balancing random segment selection, segment selection based on quality of the segment, and segment selection based on pheromone levels.

The ACO is a population based heuristic; a set of ants are allowed to individually arrive at solutions to the problem. The ants are given capabilities beyond those found in real ants. Thus, the computational ants are often called artificial ants. The primary additional capabilities are memory (to avoid duplicate segment selection) and a look-ahead capability (to select among available segments during solution construction). The ants can start at random points in the optimization problem segment space or at some common starting point; the choice of starting location varies by problem type.

Initially, ants move randomly through the network of segments. From a given segment, the ant views the remaining segments, picks one at random, and adds that segment to their partial solution. Once all ants have constructed their solution, that solution is evaluated. Naturally, there will be a combination of good and poor solutions obtained by the collective set of ants.

Once the ants have completed their solution construction, each ant deposits pheromone on the segments employed in their respective solutions. One such pheromone deposit function is:

$$\Delta\tau_{ij}^{k} = \begin{cases} \dfrac{Q}{L_k} & \text{if edge } (i,j) \text{ used} \\ 0 & \text{otherwise} \end{cases}$$

where Q is some user defined value, L_k is the solution constructed by ant k, and τ_{ij}^{K} is the pheromone deposited by ant k on segment (i, j). This pheromone change is coupled with a pheromone decay function so that a typical pheromone update function resembles:

$$\tau_{ij} = (1-\rho)\tau_{ij} + \Delta\tau_{ij}$$

$$\Delta\tau_{ij} = \sum_{k=1}^{m} \tau_{ij}^{k}.$$

The decay factor, ρ, is used to ensure the segment pheromone level, τ_{ij}, is not re-inforced too much and its loses re-inforcement over time. The value $\Delta\tau_{ij}$ is the sum of pheromone deposited by the set of ants, for which we assume, for the equation above, that there are m such ants in the set.

A final piece of the algorithm is a function that guides the ant's transition behavior. This transition function provides probability values associated with the available segments, but these values are biased by the heuristic "desirability" of the segment and the segment's current pheromone level. For instance Velayudhan et al. (2007) use

$$p_{ij}^k = s\begin{cases} \tau_{ij}^{\alpha}\nu_{ij}^{\beta} / \sum_{\ell \in s} \left[\tau_{i\ell}^{\alpha}\nu_{i\ell}^{\beta}\right] & j \in S \\ 0 & j \notin S \end{cases}$$

in the context of a traveling salesman problem application. The probability ant k, currently at node i selects segment (i, j) among their set of possible segments, $j \in S$, is p_{ij}^k. The τ_{ij} is the pheromone level on segment (i, j), the ν_{ij} is a desirability measure of the segment (such as the distance, weight or cost), and the parameters α and β are user defined and control the influence of pheromone and desirability, respectively, on the transition tendencies of the ant.

The set of ants iterate through some predetermined number of tours after which the best solution found is returned.

9.7.3 Algorithm Enhancements

The ACO has been applied to many classes of optimization problems. Velayudhan (2004) compiles a list of these applications.

A common enhancement to the ACO is to change the pheromone update strategy. Instead of allowing all ants to update, an elitist strategy allows only those ants with better problem solutions to update the pheromone trail. Another strategy is to always cause the best solution found overall to update the segments used. Both of these strategies are meant to keep pressure on those segments associated with the better solutions. Changing the timing of the update is used as well by allowing the ants to update pheromone during their tour construction vice at the completion of the tour.

Influences on the transition function can also be used. Changing the transition function form, that is changing the v_{ij} during algorithm execution, further evaporating τ_{ij}, evaporating the τ_{ij} for the segments (i, j) not used, and modifying the α and β parameters are methods that have been explored. When used, computational study is required to fine tune the changes to the problems investigated.

The ACO can quickly converge to solutions. The pressure caused by the pheromone update scheme can cause all ants to effectively construct the same solution. Too often this early convergence leads to solutions of inferior quality. A common enhancement to promote diversification in the search is to re-start the search process, add periods of random solution construction, or drastically change the α and β parameters.

9.7.4 Final Comments

The ACO is a re-inforcement learning system. The ants learn the better segments via the pheromone trail. Since the analogy is to ants, the ACO is sometimes implemented as an agent-based approach and called an agent-based optimization method. Despite being relatively new, the ACO algorithm has quickly established credibility in solving hard combinatorial optimization problems. Among its detracting features are its sometimes early convergence characteristics, the computational overhead of a population-based approach, the need for tuning the parameters of the algorithm, and its sometimes tenuous application to general IP or continuous variable optimization problems.

For further reading on the ACO, please see Dorigo et al. (1996) and the complete section on the ACO in Corne et al. (1999).

9.8 Military Applications

The range of military model applications is vast, too vast for this chapter. Battilega and Grange (1984) required an entire text to review the diverse range of models then available. In fact, their review did not even include heuristics as a major technique for military optimization applications (Battilega and Grange 1984). This section focuses on heuristic methods applied to military optimization applications.

Further, the focus is narrowed to largely the use of modern heuristic search techniques. The review here is necessarily limited in scope; we limited coverage to applications available in the open, refereed literature. A future, more comprehensive review will incorporate research products from graduate programs, analytical organizations and research laboratories, of which the applications are quite voluminous.

9.8.1 Assignment Problems

The classic assignment problem uniquely pairs objects from one set with objects from another set, while seeking that assignment that either minimizes some cost function or maximizes some profit function. Extensions add additional conditions on the assignment problem realized as constraints within the model. Military applications of the assignment problem include:

- Maintenance tasks to maintainers.
- Aircrews to aircraft.
- Personnel to organizations.
- Radio frequencies to user channels.
- Weapons to targets.
- Supplies to delivery modes.

The weapons assignment model application is crucial to military analyses, from developing requisitions to purchase the weapons to determining which mode of delivery will be used to employ the weapons. Wacholder (1989) used a neural network to solve a nonlinear weapon assignment problem for ballistic missile defense. Green et al. (1997) use a rounding heuristic to find solutions to the Arsenal Exchange Model (AEM), a model used to assign nuclear weapons to targets. Their rounding heuristic was shown to outperform the existing AEM assignment heuristic. Cullenbine et al. (2003) apply reactive tabu search to solve the Weapon Assignment Model (WAM). A WAM solution also involves the assignment of nuclear warheads to targets but those solutions are evaluated with respect to multiple goals the plan should achieve. Their solution approach included tabu search mechanisms for search diversification and was found to yield reasonable solutions faster than the currently existing heuristics. Lee et al. (2002) use an ant colony algorithm but augment their algorithm with immunity concepts inspired by Jiao and Wang (2000). Immunity concepts include vaccination and immunity processes. A vaccination process uses problem specific knowledge to modify specific aspects of the solution to avoid degradations in solution quality. An immunity process ensures weaker solutions are not used to reinforce subsequent solutions. Lee et al. (2002) compare their approach to SA and GA approaches on four randomly generated problems.

Blodgett et al. (2003) provide an interesting application of tabu search to the assignment of naval defense systems against anti-ship missiles. Rather than a simple assignment, which may require re-planning once operations commence, the authors develop a decision-tree-based solution to account for the probabilistic outcomes that arise in actual operations. The intent is to derive solutions more robust to the uncertainty of combat. Erdem and Ozdemirel (2003) apply a GA to a force-vs-force allocation problem. Their problem allocates blue (friendly) force components against red (enemy) force components to achieve defined attrition goals at minimal overall effort. The force components on each side are modeled as heterogeneous and divisible components. The GA employs problem specific chromosome repair operators to ensure feasibility of population members and it employs a local improvement phase to further improve population members (such as found in memetic algorithm approaches).

Military organizations, particularly those deployed in combat zones, require effective communications for operational success. The frequency assignment, or channel assignment, problem assigns frequencies to wireless connections to ensure communication can occur and interference from other frequencies is minimized. Aardal et al. (2007) provide a thorough survey of solution techniques for these frequency assignment problems. Their review includes classic optimization and heuristic search methods. Heuristics methods found in the literature, and covered in their survey, include greedy algorithms, local search methods, tabu search, SA, GAs, and ant colony algorithms. Lau and Tsang (2001) was not included in the survey but present a guided genetic algorithm (GGA). Their GGA is devised to retain specific knowledge of solution characteristics of the population members. This knowledge structure is used to modify the reproduction process (crossover and mutation) to re-inforce or limit strong or weak solution features, respectively. Their results indicate the GGA provides good solutions but at higher computational costs compared to local search procedures.

9.8.2 Routing Problems

A routing problem assigns a sequence of destinations to a vehicle so that some cost function is minimized as the vehicle transits the segments connecting the destinations in the vehicle's route. Military applications of routing include:

- Routing a UAV or set of UAVs over some set of targets.
- Routing a surveillance asset over some search region.
- Determining the sequence of stops for a mobility aircraft mission.
- Determining how to route a missile or aircraft to a target to avoid threats and obstacles.
- Determining how to deliver required supplies from supply locations to destination locations.

The routing problem associated with UAVs has received a great deal of attention. Additional constraints added to these problems include considerations of no-fly zones, arrival time restrictions, time-over-target restrictions, flight duration considerations, and threats within the route space. Ryan et al. (1999) examine the UAV routing problem with time windows and time precedence considerations. The authors couple a reactive tabu search route finder with a simulation route evaluator to determine "robust routes". The simulation is used to assess target coverage and travel time considerations using stochastic processes for the threats and the weather. Multiple iterations yield route segment information used to determine those better segments to use within the route. O'Rourke et al. (2001) builds upon the optimization portion of this work yielding a reactive tabu search that solves a more comprehensive routing problem. Harder et al. (2004) develop a general purpose tabu-search-based routing solver and apply the tool specifically to UAV routing. The tool consists of an interface by which the user can enter a UAV routing problem and a Java-based solver is used to return potential solutions. A subsequent empirical study involved academic benchmark problems and operationally-motivated test instances. An open software version of the solver engine is available as OpenTS. Kinney et al. (2005) consider a set of local searches, to include tabu search, for quickly solving routing problems, specifically UAV routing problems. This study included analyses regarding when to discontinue local search efforts and return the suggested solution. This quick-running heuristic approach was implemented in the Harder et al. (2004) UAV routing framework. Shetty et al. (2008) recently considered the routing of UCAVs. Their problem considers target priorities, minimum and maximum levels of target damage, payload capacities and range restrictions. They use tabu search to guide subordinate heuristics solving aspects of the decomposed problem; their tabu search guides the iterative application of an assignment and a multiple traveling salesman heuristic.

A closely related problem is that of choosing flight paths. This is particularly applicable in mission planning to avoid terrain obstacles or known threats and in search planning. John et al. (2001) use a GA to determine a flight route for a regional surveillance problem. The GA incorporates constraints for no-fly zone considerations and was able to obtain reasonable solutions in a fraction of the time required by an IP-based solver. Kierstead and DelBalzo (2003) employ a GA to design efficient search paths. While applicable to a variety of military applications, the specific application in this case is the detection of submarines within some enclosed search region.

9.8.3 Packing Problems

A packing problem seeks to place some set of objects within some specified, confined space. Attributes of these objects typically include dimensions and weight, but can expand to include myriad other attributes to include interactions with other objects requiring placement. The confined spaces can be single dimensional, like a knapsack, two-dimensional, such as floor space on a truck, or three-dimensional such as a shipping container. Typical packing applications can be found in shipping operations, military deployment load planning or storage facilities.

Packing heuristics have been widely used in mobility planning. Early approaches employed greedy algorithms. Cochard and Yost (1985) describe the Deployable Mobility Execution System, an early PC-based aircraft load planning system. The Army Automated Load Planning System (AALPS) deployed in 1978 (Heidelberg et al. 1998). Each tool provided initial load plans, using greedy construction approaches to accommodate various aircraft loading constraints and allowed the user to interactively tailor the initial load plans. Heidelberg et al. (1998) provide an improved two-dimensional packing algorithm for the gross load planning of aircraft.

Guéret et al. (2003) consider heuristics for loading aircraft. They employ two construction heuristics modified by either of two local search algorithms. The first local search approach systematically changes the order items are loaded onto the aircraft to try and improve the solution. The second local search tries to reduce the number of aircraft required by attempting to empty a selected aircraft's contents into other aircraft having available capacity. Harwig et al. (2006) apply tabu search to the two-dimensional packing problem applied to the load planning for mobility operations. This problem arises when planning shipments of material or determining airlift requirements for a planned deployment. This approach could apply to the loading of pallets or the loading of aircraft. Baltacioglu et al. (2006) develop a heuristic for three dimensional packing of cargo pallets. Their heuristic mimics how humans create packings for a confined three-dimensional space. Related work includes a tabu search approach by Lodi et al. (2002), a GA approach by Bortfeldt and Gehring (2001) and a SA approach by Faina (2000). Morgan et al. (2006) consider the problem of determining a mix of space launch vehicles for assigning a set of satellites to those vehicles. Modeled as a bin-packing problem, the authors generated over 3600 test problems, solved each to optimality or an upper bound, and examined how well a local improvement heuristic performed. The local heuristic found best solutions in 43% of the problems and on average came within 3–5% of the best known solutions for the test problems.

9.8.4 Scheduling and Planning

Scheduling and planning involves the sequencing of tasks and the assignment of resources to the tasks. As tasks are sequenced, start times, finish times, and resource expenditures are compiled and examined to ensure solution feasibility. In some applications, such as aircrew scheduling, the plethora of requirements that must be satisfied makes mathematical programming approaches impractical due to the complexity of the formulation and the time required to obtain the proven optimal solution.

Sklar et al. (1990) consider the problem of assigning crews to airlift routes to minimize the number of crews required to complete the assigned missions. They employ two constructive heuristics iteratively on decompositions of the problem. In their first heuristic, crews are assigned. These crews are used in the second heuristic to determine flight departure times. The departures are used to re-assign the crews when iterating back to the first heuristic. The process completes when the entire solution stabilizes or an iteration limit is reached. Combs and Moore (2004) apply tabu search to the crew scheduling problem. They use a hybrid approach iterating between a tabu search phase and an exact optimization phase. A set-covering problem is solved to determine crew rotations while the tabu search is executed to develop the crew schedule. The algorithm accounts for all crew rest and flight restrictions associated with the mobility air crew schedule. Lambert et al. (2007) apply tabu search to the Strategic Airlift Problem (SAP). SAP solutions:

- Pair aircraft and deployment requirements
- Determine aircraft missions required
- Time phase the aircraft along the mobility distribution network

Further details on tabu search as applied to the variety of mobility-related planning problems are found in the chapter on mobility modeling.

McGinnis et al. (1996) develop a mathematical model for scheduling training resources associated with Army initial entry training. A multi-phase heuristic is used to develop a resource schedule based on the projected trainee arrivals. The heuristic solutions are compared to exact solutions obtained via dynamic programming and are found to fall within 10% of the optimal solution, on average, but at a fraction of the computational time required. Butler et al. (1999) present an integrated system to support the scheduling of combat engineers. They employ a GA to the task of assigning the combat engineering teams to a variety of combat engineering tasks. The scheduling challenge includes assigning and sequencing combat engineering units with differing capabilities to arrive at an efficient and effective schedule to accomplish the set of tasks. Their discussion includes specifics of how the GA paradigm was tailored to the combat engineering scheduling problem.

Schlabach et al. (1999) describe FOX-GA, a GA used to support military maneuver planning. The FOX-GA implementation involves three notable contributions. First, the authors use a fast-running aggregate model to evaluate potential solutions vice using the time consuming war games that are generally used. Such solution evaluation efficiency is crucial for successful use of a GA. Second, the genetic algorithm offers the decision maker a set of solutions, vice a single solution via math programming or the full population of solutions. Finally, the authors implement niching, a GA strategy to ensure population diversity is maintained during the execution of the algorithm. Maintaining the diversity of the population ensures more efficient sampling of the solution search space.

Imsand et al. (2004) use a GA to conduct a vulnerability analysis on the Commander's Analysis and Planning Simulation (CAPS), which is used for the operational analysis of missile defense systems. The GA was used to determine those factors most influencing the simulation output, and thus affecting the planning effectiveness.

Satellite range scheduling involves the assignment of tracking stations to supported satellites to satisfy requested satellite support. Gooley et al. (1996) develop a heuristic for the Air Force Satellite Control Network scheduling problem. Their approach is to build an initial schedule and apply a local search involving interchange operators to improve the schedule. Results on an actual test problem were fair but obtained in minutes versus the longer time required by the manual process.

9.8.5 Final Comments

Our application review is by no means comprehensive; it is hopefully representative of how these heuristics are employed. We necessarily excluded from our review the vast material found in conference papers, working papers, technical reports, and graduate research documents. A scan of these materials will find the use of meta-heuristics in nearly all types of optimization applications. We also did not address the use of heuristics within a simulation-based optimization setting. Simulation-based optimization methods often use procedures based on SA, GAs, tabu search or scatter search. These methods are "wrapped" around military models and used to guide those models toward some best set of specified input parameters. For instance, Hill and McIntyre (2000) use a GA, and later a scatter-search-based technique, to find robust solutions to a multi-scenario force planning problem involving a linear mathematical program.

9.9 Concluding Remarks

A variety of trends have encouraged the development and employment of search heuristics for optimization applications. Some of these were mentioned in the preceding sections. Others are discussed in the following paragraphs.

Modern computers are tremendously powerful computational devices. Some of us still remember punch-card computer programming or the initial "personal" computers that ran just on floppy diskettes. Modern computers are fast and have lots of memory. This means we can do more calculations and search among more solutions in less time than ever before. Modern computers make heuristic search methods feasible and practical.

Modern problems are very complex. This is not to say that past problems were not complex. In fact past problems were also complex but analytical methods of the time required we simplify these complex problems to accommodate the limited analytical methods. Our modern ability to solve large problems means we sometimes no longer want to simplify reality for the sake of an algorithm. We no longer always want exact solutions to approximations of the true system of interest. We now are often more satisfied with approximate solutions to problems that are more accurate reflections of the system of interest. These more accurate problems often create intractable formulations for classical solution methods. Current decision support requirements for problem solutions make heuristic search methods preferable.

The modern decision making environment is too fast-paced to wait for a perfect answer (i.e. a guaranteed optimal answer). Getting that guaranteed optimal solution can sometimes take an inordinate amount of time. Getting an answer pretty close to an optimal can now be provided rapidly. These good, feasible solutions can be worth more to the decision maker in time-sensitive, decision-making situations. The pace of modern decision making has thus made heuristic search methods an attractive problem solving methodology.

To close, the reader should not believe we advocate exclusive use of heuristic search methods for optimization applications; we don't. When practical and appropriate, mathematical programming methods should be preferred and exploited. However, we have hopefully provided the reader with the rationale, background, and motivation to appreciate the valuable role heuristic search methods can serve in an optimization model-based decision making setting. When practical and appropriate, heuristic search methods should be explored, examined, and employed.

References

Aardal, K.I., S.P.M. van Hoesel, A.M.C.A. Koster, C. Mannino, and A. Sassano. 2007. Models and solution techniques for frequency assignment problems. *Annals of Operations Research*, 153(1), 79–129.

Baltacioglu, E., J.T. Moore, and R.R. Hill. 2006. The distributor's three dimensional pallet-packing problem: A human intelligence-based heuristic approach. *International Journal of Operational Research*, 1(3), 249–266.

Battilega, J.A., and J.K. Grange. 1984. *The Military Applications of Modeling.* Wright-Patterson Air Force Base: Air Force Institute of Technology.

Battiti, R. and G. Tecchiolli. 1994. The reactive tabu search. *ORSA Journal on Computing,* 6(2), 126–140.

Beasley, D., D.R. Bull, and R.R. Martin. 1993a. An overview of genetic algorithms 1. fundamentals. *University Computing,* 15(2), 58–69.

Beasley, D., D.R. Bull, and R.R. Martin. 1993b. An overview of genetic algorithms 2. research topics. *University Computing*, 15(4), 170–181.

Blodgett, D.E., M. Gendreau, F. Guertin, J. Potvin, and R. Séguin. 2003. A tabu search heuristic for resource management in naval warfare. *Journal of Heuristics*, 9(2), 145–169.

Bortfeldt, A. and H. Gehring. 2001. A hybrid genetic algorithm for the container loading problem. *European Journal of Operational Research*, 131(1), 143–161.

Butler, C.D., G. Boggess, and S. Bridges. 1999. Knowledge based planning of military engineering tasks. *Expert Systems with Applications*, 16(2), 221–232.

Chu, P.C. and J.E. Beasley. 1998. A genetic algorithm for the multiconstraint knapsack problem. *Journal of Heuristics*, 4, 63–86.

Cochard, D.D. and K.A. Yost. 1985. Improving utilization of air force cargo aircraft. *Interfaces*, 15(1), 53–68.

Combs, T.E. and J.T. Moore. 2004. A hybrid tabu search/set partitioning approach to tanker crew scheduling. *Military Operations Research*, 9(1), 43–56.

Corne, D., M. Dorigo, and F. Glover (Eds.). 1999. *New Ideas in Optimization*. McGraw Hill, Berkshire, England.

Cullenbine, C.A., M.A. Gallagher, and J.T. Moore. 2003. Assigning nuclear weapons with reactive tabu search. *Military Operations Research*, 8(1), 57–69.

Dorigo, M., V. Maniezzo, and A. Colorni. 1996. The ant system: Optimization by a colony of cooperating ants. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 26(1), 1–13.

Erdem, E. and N.E. Ozdemirel. 2003. An evolutionary approach for the target allocation problem. *Journal of the Operational Research Society*, 54(10), 958–969.

Faina, L. 2000. A global optimization algorithm for the three-dimensional packing problem. *European Journal of Operational Research*, 126(2), 340–354.

Glover, F. 1977. Heuristics for integer programming using surrogate constraints. *Decision Sciences*, 8(1), 156–166.

Glover, F. 1986. Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research,* 13(5), 533–549.

Glover, F. 1989. Tabu search—part i. *ORSA Journal on Computing,* 1, 190–206.

Glover, F. 1990a. Tabu search—part ii. *ORSA Journal on Computing*, 2(1), 4–32.

Glover, F. 1990b. Tabu search: A tutorial. *Interfaces*, 20(4), 74–94.

Glover, F. 1998. Tabu search—wellsprings and challenges. *European Journal of Operational Research,* 106(2–3), 221–225.

Glover, F. 2007. Tabu search—uncharted domains. *Annals of Operations Research*, 149, 89–98.

Glover, F. and M. Laguna. 1997. *Tabu Search*. Kluwer Academic Publishers, Norwell MA, USA.

Glover, F., M. Laguna, and R. Martí. 2000. Fundamentals of scatter search and path relinking. *Control and Cybernetics*, 29(3), 653–684.

Goldberg, D.E. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley. New York.

Gooley, T.D., J.J. Borsi, and J.T. Moore. 1996. Automating air force satellite control network (AFSCN) scheduling. *Mathematical and Computer Modellng*, 24(2), 91–101.

Green, D.J., J.T. Moore, and J.J. Borsi. 1997. An integer solution heuristic for the arsenal exchange model (AEM). *Military Operations Research*, 3(2), 5–15.

Guéret, C., N. Jussien, O. Lhomme, C. Pavageau, and C. Prins. 2003. Loading aircraft for military operations. *Journal of the Operational Research Society,* 54, 458–465.

Hansen, P. and N. Mladenović. 2003. Variable neighborhood search. In F. Glover and G. Kochenberger (Eds.). *Handbook of Metaheuristics*. Kluwer Academic Publishers, Norwell MA, USA.

Harder, R.W., R.R. Hill, and J.T. Moore. 2004. A java universal vehicle router for routing unmanned aerial vehicles. *International Transactions in Operational Research,* 11(3), 259–276.

Harwig, J.M., J.W. Barnes, and J.T. Moore. 2006. An adaptive tabu search approach for 2-dimensional orthogonal packing problems. *Military Operations Research,* 11(2), 5–26.

Heidelberg, K.R., G.S. Parnell, and J.E. Ames IV. 1998. Automated air load planning. *Naval Research Logistics,* 45(8), 751–768.

Henderson, D., S.H. Jacobson, and A.W. Johnson. 2003. The theory and practice of simulated annealing. In F. Glover, and G. Kochenberger (Eds.). *Handbook on Metaheuristics*. Norwell, MA: Kluwer Academic Publishers, 287–320.

Hill, R.R. and C. Hiremath. 2005. Improving genetic algorithm convergence using problem structure and domain knowledge in multidimensional knapsack problems. *International Journal of Operational Research*, 1(1/2), 145–159.

Hill, R.R. and G.A. McIntyre. 2000. A methodology for robust, multi-scenario optimization. *Phalanx,* 33(3), 27–31.

Hoff, A., A. Lokketangen, and I. Mittet. 1996. Genetic algorithms for 0/1 multidimensional knapsack problems. In *Proceedings Norsk Informatikk Konferanse, NIK '96*, Molde College, Britveien 2, 6400 Molde, Norway, 291–301.

Imsand, E.S., G. Evans, G. Dozier, and J.A. Hamilton Jr. 2004. Using genetic algorithms to aid in a vulnerability analysis of national missile defense simulation software. *Journal of Defense Modeling and Simulation*, 1(4), 215–223.

Jiao, L. and L. Wang. 2000. A novel genetic algorithm based on immunity. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 30(5), 552–561.

John, M., D. Panton, and K. White. 2001. Mission planning for regional surveillance. *Annals of Operations Research*, 108(1/4), 157–173.

Kierstead, D.P. and D.R. DelBalzo. 2003. A genetic algorithm applied to planning search paths in complicated environments. *Military Operations Research*, 8(2), 45–60.

Kinney, G.W., R.R. Hill, and J.T. Moore. 2005. Devising a quick-running heuristic for an unmanned aerial vehicle (UAV) routing systems. *Journal of the Operational Research Society*, 56(7), 776–786.

Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi. 1983. Optimization by simulated annealing. *Science,* 220(4598), 671–680.

Laguna, M. and R. Martí. 2003. *Scatter Search: Methodology and Implementations in C.* Kluwer Academic Publishers, Norwell MA, USA.

Lambert, G.R., J.W. Barnes, and D.A. Van Veldhuizen. 2007. A tabu search approach to the strategic airlift problem. *Military Operations Research*, 12(3), 59–79.

Lau, T.L. and E.P.K. Tsang. 2001. Guided genetic algorithm and its application to radio link frequency assignment problems. *Constraints*, 6(4), 373–398.

Lee, Z.J., Lee, C.Y., and S.F. Su. 2002. An immunity-based any colony optimization algorithm for solving weapon-target assignment problem. *Applied Soft Computing*, 2(1), 39–47.

Lodi, A., S. Martello, and D. Vigo. 2002. Heuristic algorithms for the three-dimensional bin packing problem. *European Journal of Operational Research*, 141(2), 410–420.

Martí, R., M. Laguna, and F. Glover. 2006. Principles of scatter search. *European Journal of Operational Research*, 169(4), 359–372.

McGinnis, M.L., E. Gernandez-Gaucherand, and P.B. Mirchandani. 1996. Military training resource scheduling: System model, optimal and heuristic decision processes. *Military Operations Research*, 2(1), 53–69.

Morgan, L.O., A.R. Morton, and R.L. Daniels. 2006. Simultaneously determining the mix of space launch vehicles and the assignment of satellites to rockets. *European Journal of Operational Research*, 17(3), 747–760.

Muller-Merbach, H. 1981. Heuristics and their design: A survey. *European Journal of Operational Research*, 8(1), 1–23.

O'Rourke, K.P., T.G. Bailey, R.R. Hill, and W.B. Carlton. 2001. Dynamic routing of unmanned aerial vehicles using reactive tabu search. *Military Operations Research*, 6(1), 5–30.

Osman, I.H. and J.P. Kelly. 1996. Meta-heuristics: An overview. In I.H. Osman and J.P. Kelly (Eds.). *Meta-Heuristics: Theory and Applications.* Kluwer Academic Publishers, Norwell MA, USA.

Reeves, C.R. 1993. *Modern Heuristics Techniques for Combinatorial Problems.* John Wiley and Sons, Inc., New York.

Reeves, C.R. 1997. Genetic algorithms for the operations researcher. *INFORMS Journal on Computing*, 9(3), 231–250.

Renner, G. and A. Ekart. 2003. Genetic algorithms in computer-aided design. *Elsevier Computer-Aided Design*, 35, 709–726.

Ross, P. 1997. What are genetic algorithms good at? *INFORMS Journal on Computing*, 9(3), 260–262.

Ryan, J.L., T.G. Bailey, and J.T. Moore. 1999. Unmanned aerial vehicle (UAV) route selection using reactive tabu search. *Military Operations Research*, 4(3), 5–24.

Schlabach, J.L., C.C. Hayes, and D.E. Goldberg. 1999. Fox-ga: A genetic algorithm for generating and analyzing battlefield courses of action. *Evolutionary Computation*, 7(1), 45–68.

Senju, S. and Y. Toyoda. 1968. An approach to linear programming with 0-1 variables. *Management Science,* 15(4), B196–B207.

Shetty, V.K., M. Sudit, and R. Nagi. 2008. Priority-based assignment and routing of a fleet of unmanned combat aerial vehicles. *Computers and Operations Research*, 35(6), 1813–1828.

Silver, E.A. 2004. An overview of heuristic solution methods. *Journal of the Operational Research Society*, 55, 936–956.

Sklar, M.G. R.D. Armstrong, and S. Samn. 1990. Heuristics for scheduling aircraft and crew during airlift operations. *Transportation Science*, 24(1), 63–76.

Velayudhan, S.P. 2004. Empirical analysis of randomness in ant colony optimization algorithms applied to the traveling salesman problem. Master's Thesis, Wright State University, Dayton, OH.

Velayudhan, S.P., R.R. Hill, C. Hiremath, and H.D. Soneji. 2007. Empirical analysis of randomness in ant colony optimization algorithms applied to the traveling salesman problem. *International Journal of Information Systems for Logistics and Management, 2*(2), 69–76.

Wacholder, E. 1989. A neural network-based optimization algorithm for the static weapon-target assignment problem. *ORSA Journal on Computing*, 1(4), 232–246.

Zalzala, M.S. and P.J. Fleming (Eds.). 1997. *Genetic Algorithms in Engineering Systems.* London: Institution of Electrical Engineers.

III

Reliability and Maintenance

10

Recent Advances in Optimal Reliability Allocation*

Way Kuo
City University of Hong Kong

Rui Wan
University of Tennessee

10.1 Introduction

Reliability has become an even greater concern in recent years because high-tech industrial processes with increasing levels of sophistication comprise most engineering systems today. Based on enhancing component reliability and providing redundancy while considering the trade-off between system performance and resources, optimal reliability design that aims to determine an optimal system-level configuration has long been an important topic in reliability engineering. Since 1960, many publications have addressed this problem using different system structures, performance measures, optimization techniques and options for reliability improvement.

Kuo and Prasad, Misra and Tillman et al. [45,93,123] provide good literature surveys of the early work in system reliability optimization. Tillman et al. [123] were the first to classify papers by system structure, problem type, and solution methods. Also described and analyzed by Tillman et al. [123] are the advantages and shortcomings of various optimization techniques. It was during the 1970s that various heuristics were developed to solve complex system reliability problems in cases where the traditional parametric optimization techniques were insufficient. In their 2000 report, Kuo and Prasad [45] summarize the developments in optimization techniques, along with recent optimization methods such as meta-heuristics, up until that time. This chapter discusses the contributions made to the literature since the publication of Kuo and Prasad [45]. The majority of recent work in this area is devoted to:

(a) Multi-state system (MSS) optimization
(b) Percentile life employed as a system performance measure
(c) Multi-objective optimization
(d) Active and cold-standby redundancy
(e) Fault-tolerance mechanism
(f) Optimization techniques, especially ant colony algorithms and hybrid optimization methods

Based on their system performance, reliability systems can be classified as binary-state systems or MSS. A binary-state system and its components may exist in only two possible states—either working or failed. Binary system reliability models have played very important roles in practical reliability engineering. To satisfactorily describe the performance of a complex system, however, we may need more than two levels of satisfaction—for example, excellent, average, and poor [46]. For this reason, multi-state system reliability models were proposed in the 1970s, although a large portion of the work devoted to MSS optimal design has emerged since 1998. The primary task of multi-state system optimization is to define the relationship between component states and system states.

Measures of system performance are basically of four kinds: (a) reliability, (b) availability, (c) mean time-to-failure, and (d) percentile life. Reliability has been widely used and thoroughly studied as the primary performance measure for nonmaintained systems. For a maintained system, however, availability, which describes the percentage of time the system really functions, should be considered instead of reliability. Availability is most commonly employed as the performance measure for renewable MSS. Meanwhile, percentile life is preferred to reliability and mean time-to-failure when the system mission time is indeterminate, as in most practical cases.

Some important design principles for improving system performance are summarized in Kuo et al. [46]. This chapter primarily reviews articles that address either the provision of redundant components in parallel or the combination of structural redundancy with the enhancement of component reliability. These are called redundancy allocation problems and reliability-redundancy allocation problems, respectively. Redundancy allocation problems are well documented elsewhere [45,46], which employ a special case of reliability-redundancy allocation problems without exploring the alternatives of component combined improvement. Recently, much of the effort in optimal reliability design has been placed on general reliability-redundancy allocation problems, rather than redundancy allocation problems.

In practice, two redundancy schemes are available: active and cold-standby. Cold-standby redundancy provides higher reliability, but it is hard to implement because of the difficulty of failure detection. Reliability design problems have generally been formulated considering active redundancy; however, an actual optimal design may include active redundancy or cold-standby redundancy or both.

However, any effort for improvement usually requires resources. Quite often it is hard for a single objective to adequately describe a real problem for which an optimal design is required. For this reason, multi-objective system design problem always deserves a lot of attention.

Optimal reliability design problems are known to be NP-hard [8]. Finding efficient optimization algorithms is always a hot spot in this field. Classification of the literature by reliability optimization techniques is summarized in Table 10.1. Meta-heuristic methods, especially genetic algorithms (GA), have been widely and successfully applied in optimal system design because of their robustness and efficiency, even though they are time consuming, especially for large problems. To improve computation efficiency, hybrid optimization algorithms have been increasingly used to achieve an effective combination of GAs with heuristic algorithms, simulation annealing methods, neural network techniques and other local search methods.

This chapter describes the state-of-art of optimal reliability design. Emphasizing the foci mentioned earlier, we classify the existing literature based on problem formulations and optimization techniques. The remainder of the chapter is organized as follows: Section 10.2 includes four main problem formulations in optimal reliability allocation; Section 10.3 describes advances related to those four types of optimization

TABLE 10.1 Reference Classification by Reliability Optimization Methods

Meta-heuristic Algorithms	
ACO	[82], [92], [95], [116]
GA	[1], [7], [11], [14], [28], [35], [52], [53], [54], [56], [57], [58], [59], [60], [62], [63], [64], [65], [67], [68], [72], [73], [74], [78], [79], [84], [91], [128]
HGA	[33], [34], [48], [49], [50], [114], [115], [121], [122], [132], [133]
TS	[41]
SA	[1], [118], [124], [131]
IA	[9]
GDA	[107]
CEA	[109]
Exact Method	
[20], [27], [30], [83], [102], [103], [119], [120]	
Max-min Approach	
[51], [104]	
Heuristic Method	
[105], [129]	
Dynamic Programming	
[97], [127]	

ACO, ant colony optimization; GA, genetic algorithm; HGA, hybrid genetic algorithm; TS, tabu search; SA. simulated annealing algorithm; IA, immune algorithm; GDA, great deluge algorithm; and CEA, cellular evolutionary approach.

problems; Section 10.4 summarizes developments in optimization techniques; and Section 10.5 provides conclusions and a discussion of future challenges related to reliability optimization problems.

10.2 Problem Formulation

Among the diversified problems in optimal reliability design, the following four basic formulations are widely covered.

10.2.1 Problem 1 (P_1)

$$\begin{aligned} &\max R_S = f(\mathbf{x}) \\ &\text{s.t.} \\ &g_i(\mathbf{x}) \le b_i, \quad \text{for } i = 1, \ldots, m \\ &\mathbf{x} \in \mathbf{X} \end{aligned}$$

where R_s, system reliability; $g_i(\mathbf{x})$, total number of resource i required for $\mathbf{x}$; b_i, upper limit of resource i; and m, number of resources.

Or

$$\begin{aligned} &\min C_S = f(\mathbf{x}) \\ &\text{s.t.} \\ &R_S \ge R_0 \\ &g_i(\mathbf{x}) \le b_i, \quad \text{for } i = 1, \ldots, m \\ &\mathbf{x} \in \mathbf{X} \end{aligned}$$

where C_s, total system cost; and R_0, a specified minimum R_s.

Problem 1 formulates the traditional reliability-redundancy allocation problem with either reliability or cost as the objective function. Its solution includes two parts: the component choices and their corresponding optimal redundancy levels.

10.2.2 Problem 2 (P_2)

$$\begin{aligned}
&\max\ t_{\alpha,\mathbf{x}} \\
&\text{s.t.} \\
&g_i(t_{\alpha,\mathbf{x}};\mathbf{x}) \le b_i, \quad \text{for } i=1,\ldots,m \\
&\mathbf{x}\in\mathbf{X}
\end{aligned}$$

where $t_{\alpha,\mathbf{x}}$, system percentile life; $t_{\alpha,\mathbf{x}}=\inf\{t\ge 0: R_S\le 1-\alpha\}$; and α, system user's risk level, $0<\alpha<1$.

Problem 2 uses percentile life as the system performance measure instead of reliability. Percentile life is preferred especially when the system mission time is indeterminate. However, it is hard to find a closed analytical form of percentile life in decision variables.

10.2.3 Problem 3 (P_3)

$$\begin{aligned}
&\max\ E(\mathbf{x},\mathbf{T},\mathbf{W}^*) \\
&\text{s.t.} \\
&g_i(\mathbf{x}) \le b_i, \quad \text{for } i=1,\ldots,m \\
&\mathbf{x}\in\mathbf{X}
\end{aligned}$$

where $\mathbf{T}$, MSS operation period; and $\mathbf{W}^*$, required MSS performance level.

Or

$$\begin{aligned}
&\min\ C_s(\mathbf{x}) \\
&\text{s.t.} \\
&E(\mathbf{x},\mathbf{T},\mathbf{W}^*) \ge E_0 \\
&g_i(\mathbf{x}) \le b_i, \quad \text{for } i=1,\ldots,m \\
&\mathbf{x}\in\mathbf{X}
\end{aligned}$$

Problem 3 represents MSS optimization problems. Here, E is used as a measure of the entire system availability to satisfy the custom demand represented by a cumulative demand curve with a known $\mathbf{T}$ and $\mathbf{W}^*$. E_0, is a specified minimum E_S and E_S, a generalized MSS availability index.

10.2.4 Problem 4 (P_4)

$$\begin{aligned}
&\max\ z=[f_1(\mathbf{x}), f_2(\mathbf{x}),\ldots, f_s(\mathbf{x})] \\
&\text{s.t.} \\
&g_i(\mathbf{x}) \le b_i, \quad \text{for } i=1,\ldots,m \\
&\mathbf{x}\in\mathbf{X}
\end{aligned}$$

For multi-objective optimization, as formulated by Problem 4, a Pareto optimal set, which includes all of the best possible trade-offs between given objectives, rather than a single optimal solution, is usually identified.

In all of the above formulations, the resource constraints may be linear or nonlinear or both.

The literature, classified by problem formulations, is summarized in Table 10.2.

TABLE 10.2 Reference Classification by Problem Formulation

P_1	[1], [2], [3], [9], [12], [13], [15], [20], [27], [30], [35], [36], [41], [48], [49], [50], [79], [81], [83], [95], [97], [100], [101], [102], [109], [119], [120], [121], [122], [124], [127], [129], [132], [133]
P_2	[11], [14], [39], [103], [132], [133]
P_3	[52], [53], [54], [56], [57], [58], [59], [60], [62], [63], [64], [65], [67], [68], [72], [73], [74], [78], [79], [84], [92], [99], [105]
P_4	[7], [16], [28], [91], [106], [114], [115], [116], [118], [131], [132]

10.3 Brief Review of Advances in P_1–P_4

10.3.1 Traditional Reliability-redundancy Allocation Problem (P_1)

System reliability can be improved either by incremental improvements in component reliability or by provision of redundancy components in parallel; both methods result in an increase in system cost. It may be advantageous to increase the component reliability to some level and provide redundancy at that level [46], i.e. the tradeoff between these two options must be considered. According to the requirements of the designers, traditional reliability-redundancy allocation problems can be formulated either to maximize system reliability under resource constraints or to minimize the total cost that satisfies the demand on system reliability. These kinds of problems have been well-developed for many different system structures, objective functions, redundancy strategies and time-to-failure distributions. Two important recent developments related to this problem are addressed below.

10.3.1.1 Active and Cold-standby Redundancy

P_1 is generally limited to active redundancy. A new optimal system configuration is obtained when active and cold-standby redundancies are both involved in the design. A cold-standby redundant component does not fail before it is put into operation by the action of switching, whereas the failure pattern of an active redundant component does not depend on whether the component is idle or in operation. Cold-standby redundancy can provide higher reliability, but it is hard to implement due to the difficulties involved in failure detection and switching.

In Coit and Liu [12], optimal solutions to reliability-redundancy allocation problems are determined for nonrepairable systems designed with multiple k-out-of-n subsystems in series. The individual subsystems may use either active or cold-standby redundancy, or they may require no redundancy. Assuming an exponentially distributed component time-to-failure with rate λ_{ij}, the failure process of subsystem i with cold-standby redundancy can be described by a Poisson process with rate $\lambda_{ij}k_i$, while the subsystem reliability with active redundancy is computed by standard binominal techniques.

For series-parallel systems with only cold-standby redundancy, Coit [13] employs the more flexible and realistic Erlang distributed component time-to-failure. Subsystem reliability can still be evaluated through a Poisson process, though $\rho_i(t)$ must be introduced to describe the reliability of the imperfect detection/switching mechanism for each subsystem.

Coit [15] directly extends this earlier work by introducing the choice of redundancy strategies as an additional decision variable. With imperfect switching, it illustrates that there is a maximum redundancy level where cold-standby reliability is greater than, or equal to, active reliability, i.e. cold-standby redundancy is preferable before this maximum level while active redundancy is preferable after that.

All three problems formulated above can be transformed by logarithm transformation and by defining new 0–1 decision variables. This transformation linearizes the problems and allows for the use of integer programming algorithms. For each of these methods, however, no mixture of component types or redundancy strategies is allowed within any of the subsystems.

In addition, Bueno [6] investigates the problem of where to allocate a spare in a k-out-of-n: F system of dependent components through minimal standby redundancy; and Romera et al. [110] studies the allocation of one active redundancy when it differs based on the component with which it is to be allocated. Prasad et al. [101] considers the problem of optimally allocating a fixed number of s-identical multifunctional spares for a deterministic or stochastic mission time. In spite of some sufficiency conditions for optimality, the proposed algorithm can be easily implemented even for large systems.

10.3.1.2 Fault-tolerance Mechanism

Fault-tolerance is the ability of a system to continue performing its intended function in spite of faults. System designs with fault-tolerance mechanisms are particularly important for some computer-based systems with life-critical applications, since they must behave like a nonrepairable system within each mission, and maintenance activities are performed only when the system is idle [3].

Amari et al. [2] maximizes the reliability of systems subjected to imperfect fault-coverage. It generalizes that the reliability of such a system decreases with an increase in redundancy after a particular limit. The results include the effect of common-cause failures and the maximum allowable spare limit. The models considered include parallel, parallel-series, series-parallel, k-out-of-n and k-out-of-$(2k-1)$ systems.

Similarly to imperfect fault-coverage, Amari et al. [3] later assumes the redundancy configurations of all subsystems in a nonseries-parallel system are fixed except the k-out-of-n: G subsystem being analyzed. The analysis leads to n^*, the optimal number of components maximizing the reliability of this subsystem, which is shown to be necessarily greater than, or equal to, the optimal number required to maximize the reliability of the entire system. It also proves that n^* offers exactly the maximal system reliability if the subsystem being analyzed is in series with the rest of the system. These results can even be extended to cost minimization problems.

Lyu et al. [87] considers software component testing resource allocation for a system with single or multiple applications, each with a prespecified reliability requirement. Given the coverage factors, it can also include fault-tolerance mechanisms in the problem formulation. The relationship between the component failure rates of and the cost of decreasing this rate is modeled by various types of reliability-growth curves.

For software systems, Levitin [79] presents an universal generating function (UGF) and GA based algorithm that selects the set of versions and determines the sequence of their execution, such that the system reliability (defined as the probability of obtaining the correct output within a specified time) is maximized subject to cost constraints. The software system is built from fault-tolerant N-version programming (*NVP*) and recovery block (*RB*) components.

All of these optimization models mentioned above have been developed for hardware-only or software-only systems. Wattanapongsakorn and Levitan [124] first considers several simple configurations of fault-tolerant embedded systems (hardware and software, *RB/i/j*, recovery block architecture that can tolerate i hardware and j software faults; *NVP/i/j*, N-version programming architecture that can tolerate i hardware and j software faults) including *NVP/0/1, NVP/1/1,* and *RB/1/1*, where failures of software units are not necessarily statistically independent. A real-time embedded system is used to demonstrate and validate the models solved by a simulated annealing optimization algorithm. Moreover, Levitin [80] generally takes into account fault-tolerant systems with series architecture and arbitrary number of hardware and software versions without common cause failures. An important advantage of the presented algorithm lies in its ability to evaluate both system reliability and performance indices.

10.3.2 Percentile Life Optimization Problem (P_2)

Many diversified models and solution methods, where reliability is used as the system performance measure, have been proposed and developed since the 1960s. However, this is not an appropriate choice when mission time cannot be clearly specified or a system is intended for use as long as it functions. Average life is also not reliable, especially when the implications of failure are critical or the variance in the system life is high. Percentile life is considered to be a more appropriate measure, since it incorporates system designer and user risk. When using percentile life as the objective function, the main difficulty is its mathematical inconvenience, because it is hard to find a closed analytical form of percentile life in the decision variables.

Coit and Smith [11] solves redundancy allocation problems for series-parallel systems where the objective is to maximize a lower percentile of the system time-to-failure (*TTF*) distribution. Component *TTF* has a Weibull distribution with known deterministic parameters. The proposed algorithm uses a genetic algorithm to search the prospective solution-space and a bisection search to evaluate t' in $R(t', \mathbf{x}{:}\mathbf{k}) = 1 - \alpha$. It is demonstrated that the solution that maximizes the reliability is not particularly effective at maximizing system percentile life at any α level, and the recommended design configurations are very different depending on the α level. Later in the literature, Coit and Smith [14] addresses similar problems where Weibull shape parameters are accurately estimated but scale parameters are random variables following a uniform distribution.

Prasad et al. [103] develops a lexicographic search methodology that is the first to provide exact optimal redundancy allocations for percentile life optimization problems. The continuous relaxed problem, solved by Kuhn-Tucker conditions and a two-stage hierarchical search, is considered for obtaining an upper bound, which is used iteratively to effectively reduce search space. This algorithm is general for any continuous increasing lifetime distribution.

Three important results are presented in Kim and Kuo [39] which describe the general relationships between reliability and percentile life maximizing problems.

(a) S_2 equals S_1 given $\alpha_t = 1 - R_s(t, \mathbf{x}_t^*)$, where S_1 is the set of optimal solutions of P_1 and $\mathbf{x}_t^* \in S_1$.
(b) S_1 equals S_2 given t_{α, x_α^*}, where $\mathbf{x}_\alpha^* \in S_2$.
(c) Let $\psi(t)$ be the optimal objective value of P_1. For a fixed α, $t_\alpha = \inf\{t \geq 0 : \Psi(t) \leq 1 - \alpha\}$ is the optimal objective value of P_2.

Based on these results, a methodology for P_2 is proposed to repeat solving P_1 under different mission times satisfying $\psi(t_0) \geq 1 - \alpha \geq \psi(t_2)$ until $\mathbf{x}_{t=t_0}^* = \mathbf{x}_{t=t_2}^*$ and $t_0 - t_2$ is within a specified tolerance. It is reported to be capable of settling many unsolved P_2s using existing reliability-optimization algorithms. Without the necessity of an initial guess, this method is much better than [103] at reaching the exact optimal solution in terms of execution time.

10.3.3 MSS Optimization (P_3)

MSS is defined as a system that can unambiguously perform its task at different performance levels, depending on the state of its components which can be characterized by nominal performance rate, availability and cost. Based on their physical nature, multi-state systems can be classified into two important types: Type I MSS (e.g. power systems) and Type II MSS (e.g. computer systems), which use capacity and operation time as their performance measures, respectively.

In the literature, an important optimization strategy, combining UGF and GA, has been well developed and widely applied to reliability optimization problems of renewable MSS. In this strategy, there are two main tasks:

(a) According to the system structure and the system physical nature, obtain the system UGF from the component UGFs.
(b) Find an effective decoding and encoding technique to improve the efficiency of the GA.

Levitin et al. [52] first uses an UGF approach to evaluate the availability of a series-parallel multi-state system with relatively small computational resources. The essential property of the *U*-transform enables the total *U*-function for a MSS with components connected in parallel, or in series, to be obtained by simple algebraic operations involving individual component *U*-functions. The operator Ω_ω is defined by Equations 10.1 through 10.3.

$$\Omega_\omega(U_1(z),U_2(z)) = \Omega_\omega\left[\sum_{i=1}^{I} p_{1i} z^{g_{1i}}, \sum_{j=1}^{J} p_{2j} z^{g_{2j}}\right] = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{1i} p_{2j} z^{\omega(g_{1i},g_{2j})} \quad (10.1)$$

$$\Omega_\omega(U_1(z),\cdots,U_k(z),U_{k+1}(z),\cdots U_n(z)) = \Omega_\omega(U_1(z),\cdots,U_{k+1}(z),U_k(z),\cdots U_n(z)) \quad (10.2)$$

$$\Omega_\omega(U_1(z),\cdots,U_k(z),U_{k+1}(z),\cdots U_n(z)) = \Omega_\omega(\Omega_\omega(U_1(z),\cdots,U_k(z)),\Omega_\omega(U_{k+1}(z),\cdots U_n(z))) \quad (10.3)$$

The function $\omega(\cdot)$ takes the form from Equations 10.4 through 10.7.
For Type I MSS,

$$\omega_{s1}(g_1,g_2) = \min(g_1,g_2) \quad (10.4)$$

$$\omega_{p1}(g_1,g_2) = g_1 + g_2 \quad (10.5)$$

For Type II MSS,

$$\omega_{s2}(g_1,g_2) = \frac{g_1 * g_2}{g_1 + g_2} \quad (10.6)$$

$$\omega_{p2}(g_1,g_2) = g_1 + g_2 \quad (10.7)$$

Later, Levitin and Lisnianski [55] combines importance and sensitivity analysis and Levitin [75] extends this UGF approach to MSS with dependent elements. Table 10.3 summarizes the application of UGF to some typical MSS structures in optimal reliability design.

With this UGF and GA strategy, Lisnianski et al. [84] solves the structure optimization of a multi-state system with time redundancy. TRS can be treated as a Type II MSS, where the system and its component performance are measured by the processing speed. Two kinds of systems are considered: systems with hot reserves and systems with work sharing between components connected in parallel.

TABLE 10.3 Application of UGF Approach

Series-parallel system	[52], [54], [55], [59], [62], [65], [72], [73], [74], [92]
Bridge system	[53], [56], [71], [84]
Linear multi-state sliding-window system	[64], [70], [78]
Weighted voting system	[58], [66]
Acyclic transmission network	[63], [69]
Linear multi-state consecutively connected system	[67], [68]
Acyclic consecutively connected network	[61]

Levitin [57] applies the UGF and GA strategy to a multi-state system consisting of two parts:

(a) RGS including a number of resource generating subsystems.
(b) MPS including elements that consume a fixed amount of resources to perform their tasks.

Total system performance depends on the state of each subsystem in the RGS and the maximum possible productivity of the MPS. The maximum possible productivity of the MPS is determined by an integer linear programming problem related to the states of the RGS.

Levitin and Lisnianski [59] develops an UGF and GA strategy for multi-state series-parallel systems with two failure modes: open mode and closed mode. Two optimal designs are found to maximize either the system availability or the proximity of expected system performance to the desired levels for both modes. The function $\omega(\cdot)$ and the conditions of system success for both two modes are shown as follows.

For Type I MSS:

$$\omega_s^O(g_1, g_2) = \omega_s^C(g_1, g_2) = \min(g_1, g_2) \tag{10.8}$$

$$\omega_p^O(g_1, g_2) = \omega_p^C(g_1, g_2) = g_1 + g_2 \tag{10.9}$$

$$F_C(G_C, W_C) = G_C - W_C \geq 0 \tag{10.10}$$

$$F_O(G_O, W_O) = W_O - G_O \geq 0 \tag{10.11}$$

For Type II MSS:

$$\omega_s^O(g_1, g_2) = \min(g_1, g_2) \tag{10.12}$$

$$\omega_p^O(g_1, g_2) = \max(g_1, g_2) \tag{10.13}$$

$$\omega_s^C(g_1, g_2) = \max(g_1, g_2) \tag{10.14}$$

$$\omega_p^C(g_1, g_2) = \min(g_1, g_2) \tag{10.15}$$

$$F_C(G_C, W_C) = W_C - G_C \geq 0 \tag{10.16}$$

$$F_O(G_O, W_O) = W_O - G_O \geq 0 \tag{10.17}$$

Thus, the system availability can be denoted by

$$A_s(t) = 1 - \Pr\{F_C(G_C(t), W_C) < 0\} - \Pr\{F_O(G_O(t), W_O) < 0\} \tag{10.18}$$

Later Levitin [65,71] introduces a probability parameter of 0.5 for both modes and even extends this technique to evaluate the availability of systems with bridge structures [71].

To describe the ability of a multi-state system to tolerate both internal failures and external attacks, survivability, instead of reliability, is proposed [56,60,67,72–74]. Levitin and Lisnianski [56] considers the problem of how to separate the elements of the same functionality between two parallel bridge components in order to achieve the maximal level of system survivability, while an UGF and GA strategy in Levitin and Lisnianski [60] is used to solve the more general survivability optimization problem of how to separate the elements of a series-parallel system under the constraint of a separation cost. Levitin, and Lisnianski [72] considers the problem of finding structure of series-parallel multi-state system (including choice of system elements, their separation and protection) in order to achieve a desired level of system survivability by minimal cost. To improve system's survivability, Levitin [73,74] further applies a multi-level protection to its subsystems and the choice of structure of multilevel protection and choice of protection methods are also included. Other than series-parallel system, Levitin [67] provides the optimal allocation of multistate LCCS with vulnerable nodes. It should be noted that the solution that provides the maximal system survivability for a given demand does not necessarily provide the greatest system expected performance rate and that the optimal solutions may be different when the system operates under different vulnerabilities.

In addition to a GA, Massim et al. [92] presents an ant colony method that combines with a UGF technique to find an optimal series-parallel power structure configuration.

Besides this primary UGF approach, a few other methods have been proposed for MSS reliability optimization problems. Ramirez-Marquez and Coit [105] develops a heuristic algorithm RAMC for a Type I multi-state series-parallel system. The availability of each subsystem is determined by a binomial technique, and, thus, the system availability can be obtained in a straightforward manner from the product of all subsystem availabilities without using UGF. Nevertheless, this algorithm can only adapt to relatively simple formulations, including those with only two-state component behavior and no mixing of functionally equivalent components within a particular subsystem.

A novel continuous-state system model, which may represent reality more accurately than a discrete-state system model, is presented in Liu et al. [86]. Given the system utility function and the component state probability density functions, a neural network approach is developed to approximate the objective reliability function of this continuous MSS optimal design problem.

10.3.4 Multi-objective Optimization (P_4)

In the previous discussion, all problems were single-objective. Rarely does a single objective with several hard constraints adequately represent a real problem for which an optimal design is required. When designing a reliable system, as formulated by P_4, it is always desirable to simultaneously optimize several opposing design objectives such as reliability, cost, even volume and weight. For this reason, a recently proposed multiobjective system design problem deserves a lot of attention. The objectives of this problem are to maximize the system reliability estimates and minimize their associated variance while considering the uncertainty of the component reliability estimations. A Pareto optimal set, which includes all of the best possible trade-offs between the given objectives, rather than a single optimal solution, is usually identified for multiobjective optimization problems.

When considering complex systems, the reliability optimization problem has been modeled as a fuzzy multi-objective optimization problem in Ravi et al. [106], where linear membership functions are used for all of the fuzzy goals. With the Bellman and Zadeh model, the decision is defined as the intersection of all of the fuzzy sets represented by the objectives.

$$\mu_{\tilde{D}}(x) = (\mu_{\tilde{f}_1}(x) * \cdots * \mu_{\tilde{f}_m}(x)) \tag{10.19}$$

The influence of various kinds of aggregators, such as the *product* operator, *min* operator, *arithmetic mean* operator, fuzzy, and a convex combination of the *min* and the *max* operator and γ-operator on the solution is also studied primarily to learn each's advantage over the noncompensatory *min* operator. It was found that in some problems illustrated in this paper, the fuzzy *and* the convex combination of the *min* and the *max* operator yield efficient solutions.

Sasaki and Gen [114,115] solve multiobjective reliability-redundancy allocation problems using similar linear membership functions for both objectives and constraints. By introducing 0–1 variables and by using an *add* operator to obtain the weighted sum of all the membership functions, the problem is transformed into a bi-criteria single-objective linear programming problem with generalized upper bounding (GUB) constraints. The proposed hybrid GA makes use of the GUB structure and combines it with a heuristic approach to improve the quality of the solutions at each generation.

With a weighting technique, Elegbede and Adjallah [28] also transfers P_4 into a single-objective optimization problem and proposes a GA-based approach whose parameters can be adjusted with the experimental plan technique. Busacca et al. [7] develops a multi-objective GA to obtain an optimal system configuration and inspection policy by considering every target as a separate objective. Both problems have two objectives: maximization of the system reliability and minimization of the total cost, subject to resource constraints.

P_4 is considered for series-parallel systems, *R*/*B*/1/1, and bridge systems in Coit et al. [16] with multiple objectives to maximize the system reliability while minimizing its associated variance when the component reliability estimates are treated as random variables. For series-parallel systems, component reliabilities of the same type are considered to be dependent since they usually share the same reliability estimate from a pooled data set. The system variance is straightforwardly expressed as a function in the higher moments of the component unreliability estimates [38]. For *R*/*B*/1/1, the hardware components are considered identical and statistically independent, while even independently developed software versions are found to have related faults as presented by the parameters *Prv* and *Pall* (*Pry*, probability of failure from related fault between two software versions; *Pall*, probability of failure from related fault among all software versions due to faults in specification). Pareto optimal solutions are found by solving a series of weighted objective problems with incrementally varied weights. It is worth noting that significantly different designs are obtained when the formulation incorporates estimation uncertainty or when the component reliability estimates are treated as statistically dependent. Similarly, Marseguerra et al. [91] utilizes a multi-objective GA to select an optimal network design that balances the dual objectives of high reliability and low uncertainty in its estimation. But the latter exploits Monte Carlo simulation as the objective function evaluation engine.

Zafiropoulos and Dialynas [131] presents an efficient computational methodology to obtain the optimal system structure of a static transfer switch, a typical power electronic device. This device can be decomposed into several components and its equivalent reliability block diagram is obtained by the minimal cut set method. Because of the existence of unit-to-unit variability, each component chosen from several off-the-shelf types is considered with failure rate uncertainty, which is modeled by a normal, or triangular, distribution. The simulation of the component failure rate distributions is performed using the Latin Hypercube Sampling method, and a simulated annealing algorithm (SA) is finally applied to generate the Pareto optimal solutions.

Shelokar et al. [116] illustrates the application of the ant colony optimization algorithm to solve both continuous function and combinatorial optimization problems in reliability engineering. The single or multiobjective reliability optimization problem is analogized to Dorigo's TSP problem, and a combinatorial algorithm, which includes a global search inspired by a GA coupled with a pheromone-mediated local search, is proposed. After the global search, the pheromone values for the newly created solution are calculated by a weighted average of the pheromone values of the corresponding parent solutions. A trial solution for conducting the local search is selected with a probability proportional to its current pheromone trial value. A two-step strength Pareto fitness assignment procedure is combined to handle multi-objective problems. The advantage of employing the ant colony heuristic for multiobjective problems is that it can produce the entire set of optimal solutions in a single run.

Suman [118] tests five simulated annealing-based multiobjective algorithms—suppapitnarm multiobjective simulated annealing (SMOSA); ulungu multi-objective simulated annealing (UMOSA); Pareto simulated annealing (PSA); Pareto domination based multi-objective simulated annealing (PDMOSA) and weight based multiobjective simulated annealing (WMOSA). Evaluated by 10 comparisons, measure *C* is introduced to gauge the coverage of two approximations for the real nondominated set. From the analysis, the computational cost of the WMOSA is the lowest, and it works well even when a large number of constraints are involved, while the PDMOSA consumes more computational time and may not perform very well for problems with too many variables.

10.4 Developments in Optimization Techniques

This section reviews recent developments of heuristic algorithms, meta-heuristic algorithms, exact methods and other optimization techniques in optimal reliability design. Due to their robustness and feasibility, meta-heuristic algorithms, especially GAs, have been widely and successfully applied. To improve computation efficiency or to avoid premature convergence, an important part of this work has been devoted in recent years to developing hybrid genetic algorithms, which usually combine a GA with heuristic algorithms, simulation annealing methods, neural network techniques or other local search methods. Though more computation effort is involved, exact methods are particularly advantageous for small problems, and their solutions can be used to measure the performance of the heuristic or meta-heuristic methods [45]. No obviously superior heuristic method has been proposed, but several of them have been well combined with exact or meta-heuristic methods to improve their computation efficiency.

10.4.1 Meta-heuristic Methods

Meta-heuristic methods inspired by natural phenomena usually include the genetic algorithm, tabu search, simulated annealing algorithm and ant colony optimization method. Ant colony optimization (ACO) has been recently introduced into optimal reliability design, and it is proving to be a very promising general method in this field. Altiparmak [1] provides a comparison of meta-heuristics for the optimal design of computer networks.

10.4.1.1 Ant Colony Optimization Method

ACO is one of the adaptive meta-heuristic optimization methods developed by Dorigo [21] for traveling salesman problems and further improved by him [22–26]. It is inspired by the behavior of real life ants that consistently establish the shortest path from their nest to food. The essential trait of the ACO algorithm is the combination of *a priori* information about the structure of a promising solution with *posteriori* information about the structure of previously obtained good solutions [88].

Liang and Smith [81] first develops an ant colony meta-heuristic optimization method to solve the reliability-redundancy allocation problem for a k-out-of-n: G series system. The proposed ACO approach includes four stages:

(a) Construction stage: construct an initial solution by selecting component j for subsystem i according to its specific heuristic η_{ij} and pheromone trail intensity τ_{ij}, which also sets up the transition probability mass function P_{ij}.
(b) Evaluation stage: evaluate the corresponding system reliability and penalized system reliability providing the specified penalized parameter.
(c) Improvement stage: improve the constructed solutions through local search.
(d) Updating stage: update the pheromone value online and offline given the corresponding penalized system reliability and the controlling parameter for pheromone persistence.

Nahas and Nourelfath [95] presents an application of the ant system in a reliability optimization problem for a series system, with multichoice constraints incorporated at each subsystem, to maximize the system reliability subject to the system budget. It also combines a local search algorithm and a specific improvement algorithm that uses the remaining budget to improve the quality of a solution.

The ACO algorithm has also been applied to a multiobjective reliability optimization problem [116] and to the optimal design of multistate series-parallel power systems [92].

10.4.1.2 Hybrid Genetic Algorithm

GA is a population-based directed random search technique inspired by the principles of evolution. Though it provides only heuristic solutions, it can be effectively applied to almost all complex combinatorial problems, and, thus, it has been employed in a large number of references as shown in Table 10.1. Gen and Kim [29] provides a state-of-the-art survey of GA-based reliability design.

To improve computational efficiency, or to avoid premature convergence, numerous researchers have been inspired to seek effective combinations of GAs with heuristic algorithms, simulation annealing methods, neural network techniques, steepest decent methods or other local search methods. The combinations are generally called hybrid genetic algorithms, and they represent one of the most promising developmental directions in optimization techniques.

Considering a complex system with a known system structure function, Zhao and Liu [132] provides a unified modeling idea for both active and cold-standby redundancy optimization problems. The model prohibits any mixture of component types within subsystems. Both the lifetime and the cost of redundancy components are considered as random variables, so stochastic simulation is used to estimate the system performance, including the mean lifetime, percentile lifetime and reliability. To speed up the solution process, these simulation results become the training data for training a neural network to approximate the system performance. The trained neural network is finally embedded into a genetic algorithm to form a hybrid intelligent algorithm for solving the proposed model. Later, Zhao and Liu [133] uses random fuzzy lifetimes as the basic parameters and employs a random fuzzy simulation to generate the training data.

Lee et al. [48] develops a two-phase neural network-hybrid genetic algorithm (NN-hGA) in which NN is used as a rough search technique to devise the initial solutions for a GA. By bounding the broad continuous search space with the NN technique, the NN-hGA derives the optimum robustly. However, in some cases, this algorithm may require too much computational time to be practical.

To improve the computation efficiency, Lee et al. [49] presents a NN-flcGA to effectively control the balance between exploitation and exploration which characterizes the behavior of GAs. The essential features of the NN-flcGA include:

(a) Combination with a NN technique to devise initial values for the GA
(b) Application of a fuzzy logic controller when tuning strategy GA parameters dynamically
(c) Incorporation of the revised simplex search method

Later, Lee et al. [50] proposes a similar hybrid GA called f-hGA for the redundancy allocation problem of a series-parallel system. It is based on

(a) Application of a fuzzy logic controller to automatically regulate the GA parameters; and
(b) Incorporation of the iterative hill climbing method to perform local exploitation around the near optimum solution

Hsieh and Hsieh [33] considers the optimal task allocation strategy and hardware redundancy level for a cycle-free distributed computing system so that the system cost during the period of task execution is minimized. The proposed hybrid heuristic combines the GA and the steepest decent method. Later Hsieh [34] seeks similar optimal solutions to minimize system cost under constraints on the hardware redundancy levels. Based on the GA and a local search procedure, a hybrid GA is developed and compared with the simple GA. The simulation results show that the hybrid GA provides higher solution quality with less computational time.

10.4.1.3 Tabu Search

Though Kuo and Prasad [45] describes the promise of tabu search (TS), Kulturel-Konak et al. [41] first develops a TS approach with the application of near feasible threshold (NFT) [10] for reliability optimization problems. This method uses a subsystem-based tabu entry and dynamic length tabu list to reduce the sensitivity of the algorithm to selection of the tabu list length. The definition of the moves in this approach offers an advantage in efficiency, since it does not require recalculating the entire system reliability, but only the reliability of the changed subsystem. The results of several examples demonstrate the superior performance of this TS approach in terms of efficiency and solution superiority when compared to that of a GA.

10.4.1.4 Other Meta-heuristic Methods

Some other adaptive meta-heuristic optimization methods inspired by activities in nature have also been proposed and applied in optimal reliability design. Chen and You [9] develops an immune algorithms-based approach inspired by the natural immune system of all animals. It analogizes antibodies and antigens as the solutions and objection functions, respectively. Rocco [109] proposes a cellular evolutionary approach (CEA) combining the multimember evolution strategy with concepts from Cellular Automata [125] for the selection step. In this approach, the parents' selection is performed only in the neighborhood in contrast to the general evolutionary strategy that searches for parents in the whole population. And a great deluge algorithm (GDA) is extended and applied to optimize the reliability of complex systems by Ravi [107]. When both accuracy and speed are considered simultaneously, it is proven to be an efficient alternative to ACO and other existing optimization techniques.

10.4.2 Exact Methods

Unlike meta-heuristic algorithms, exact methods provide exact optimal solutions though much more computation complexity is involved. The development of exact methods, such as the branch-and-bound approach and lexicographic search, has recently been concentrated on techniques to reduce the search space of discrete optimization methods.

Sung and Cho [119] considers a reliability-redundancy allocation problem in which multiple-choice and resource constraints are incorporated. The problem is first transformed into a bi-criteria nonlinear integer programming problem by introducing 0–1 variables. Given a good feasible solution, the lower reliability bound of a subsystem is determined by the product of the maximal component reliabilities of all the other subsystems in the solution, while the upper bound is determined by the maximal amount of available sources of this subsystem. A branch-and-bound procedure, based on this reduced solution space, is then derived to search for the global optimal solution. Later, Sung and Cho [120] even combines the lower and upper bounds of the system reliability, which are obtained by variable relaxation and Lagrangean relaxation techniques, to further reduce the search space.

Also with a branch-and-bound algorithm, Djerdjour and Rekab [20] obtains the upper bound of a series-parallel system reliability from its continuous relaxation problem. The relaxed problem is efficiently solved by the greedy procedure described by Djerdjour [19], combining heuristic methods to make use of some slack in the constraints obtained from rounding down. This technique assumes the objective and constraint functions are monotonically increasing. Ha and Kuo [30] presents an efficient branch-and-bound approach for coherent systems based on a 1-neighborhood local maximum obtained from the steepest ascent heuristic method. Numerical examples of a bridge system and a hierarchical series-parallel system demonstrate the advantages of this proposed algorithm in flexibility and efficiency.

Apart from the branch-and-bound approach, Prasad and Kuo [102] presents a partial enumeration method for a wide range of complex optimization problems based on a lexicographic search. The proposed upper bound of system reliability is very useful in eliminating several inferior feasible or infeasible solutions as shown in either big or small numerical examples. It also shows that the search process described in Nahas and Nourelfath [95] does not necessarily give an exact optimal solution due to its logical flows.

Lin and Kuo [83] develop a strong heuristic to search for an ideal allocation through the application of the reliability importance. It concludes that, if there exists an invariant optimal allocation for a system, the optimal allocation is to assign component reliabilities according to B-importance ordering. This Lin and Kuo heuristic can provide an exact optimal allocation.

Assuming the existence of a convex and differential reliability cost function $C_i(y_{ij})$, $y_{ij} = \log(1 - r_{ij})$ for all component j in any subsystem i, Elegbede et al. [27] proves that the components in each subsystem of a series-parallel system must have identical reliability for the purpose of cost minimization. The solution of the corresponding unconstrained problem provides the upper bound of the cost, while a doubly minimization problem gives its lower bound. With these results, the algorithm ECAY, which can provide either exact or approximate solutions depending on different stop criteria, is proposed for series-parallel systems.

10.4.3 Other Optimization Techniques

For series-parallel systems, Ramirez-Marquez et al. [104] formulates the reliability-redundancy optimization problem with the objective of maximizing the minimal subsystem reliability.

10.4.4 Problem 5 (P_5)

$$\max_{x_1, x_2, \ldots} (\min_i (1 - \prod_j (1 - r_{ij})^{x_{ij}}))$$

$$\text{s.t.}$$

$$g_i(\mathbf{x}) \le b_i, \text{ for } i = 1, \ldots m$$

$$\mathbf{x} \in \mathbf{X}$$

Assuming linear constraints, an equivalent linear formulation of P_5 [36] can be obtained through an easy logarithm transformation, and, thus, the problem can be solved by readily available commercial software. It can also serve as a surrogate for traditional reliability optimization problems accomplished by sequentially solving a series of max-min subproblems.

Lee et al. [51] presents a comparison between the Nakagawa and Nakashima method [43] and the max-min approach used by Ramirez-Marquez from the standpoint of solution quality and computational complexity. The experimental results show that the max-min approach is superior to the Nakagawa and Nakashima method in terms of solution quality in small-scale problems, but the analysis of its computational complexity demonstrates that the max-min approach is inferior to other greedy heuristics.

You and Chen [129] develops a heuristic approach inspired by the greedy method and a GA. The structure of this algorithm includes:

(a) randomly generating a specified population size number of minimum workable solutions.
(b) assigning components either according to the greedy method or to the random selection method.
(c) improving solutions through an inner-system and inter-system solution revision process.

Ng and Sancho [97] applies a hybrid dynamic programming/depth-first search algorithm to redundancy allocation problems with more than one constraint. Given the tightest upper bound, the knapsack relaxation problem is formulated with only one constraint, and its solution $f_1(b)$ is obtained by a dynamic programming method. After choosing a small specified parameter e, the depth-first search technique is used to find all near-optimal solutions with objectives between $f_1(b)$ and $f_1(b)-e$. The optimal solution is given by the best feasible solution among all of the near-optimal solutions.

Yalaoui et al. [127] also presents a new dynamic programming method for a reliability-redundancy allocation problem in series-parallel systems where components must be chosen among a finite set. This pseudopolynomial YCC algorithm is composed of two steps: the solution of the sub-problems, one for each subsystem, and the global resolution using previous results. It shows that the solutions converge quickly toward the optimum as a function of the required precision.

10.5 Comparisons and Discussions of Algorithms Reported in Literature

In this section, we provide a comparison of several heuristic or meta-heuristic algorithms reported in the literature. The compared numerical results are from the GA in Coit and Smith [10], the ACO in Liang and Smith [81], TS in Kulturel-Konak et al. [41], linear approximation in Hsieh [36], the immune algorithm (IA) in Chen and You [9] and the heuristic method in You and Chen [129]. The 33 variations of the Fyffe et al. problem, as devised by Nakagawa and Miyazaki [96], are used to test their performance, where different types are allowed to reside in parallel. In this problem set, the cost constraint is maintained at 130 and the weight constraint varies from 191 to 159.

As shown in Table 10.4, ACO [81], TS [41], IA [36] and heuristic methods [129] generally yield solutions with a higher reliability. When compared to GA [10],

(a) ACO [81] is reported to consistently perform well over different problem sizes and parameters and improve on GA's random behavior.
(b) TS [41] results in a superior performance in terms of best solutions found and reduced variability and greater efficiency based on the number of objective function evaluations required.
(c) IA [9] finds better or equally good solutions for all 33 test problems, but the performance of this IA-based approach is sensitive to value-combinations of the parameters, whose best values are case-dependent and only based upon the experience from preliminary runs. And more CPU time is taken by IAs.
(d) The best solutions found by heuristic methods [129] are all better than, or as good as, the well-known best solutions from other approaches. With this method, the average CPU time for each problem is within 8 seconds.
(e) In terms of solution quality, the proposed linear approximation approach [36] is inferior. But it is very efficient and the CPU time for all of the test problems is within one second.
(f) If a decision-maker is considering the max-min approach as a surrogate for system reliability maximization, the max-min approach [104] is shown to be capable of obtaining a close solution (within 0.22%), but it is unknown whether this performance will continue as problem sizes become larger.

For all the optimization techniques mentioned above, it might be hard to discuss about which tool is superior because in different design problems or even in a same problem with different parameters, these tools will perform variously.

Generally, if computational efficiency is of most concern to designer, linear approximation or heuristic methods can obtain competitive feasible solutions within a very short time (few seconds), as reported [36,129]. The proposed linear approximation [36] is also easy to implement with any LP software. But the main limitation of those reported approaches is that the constraints must be linear and separable.

Due to their robustness and feasibility, meta-heuristic methods such as GA and recently developed TS and ACO could be successfully applied to almost all NP-hard reliability optimization problems. However, they can not guarantee the optimality and sometimes can suffer from the premature convergence situation of their solutions because they have many unknown parameters and they neither use a

TABLE 10.4 Comparison of Several Algorithms in the Literature Each for the Test Problem

	System Reliability					
Weight	GA [10]	ACO [81]	TS [41]	Hsieh [36]	IA [9]	You and Chen [129]
191	0.98670	0.9868	0.98681	0.98671	0.98681	0.98681
190	0.98570	0.9859	0.98642	0.98632	0.98642	0.98642
189	0.98560	0.9858	0.98592	0.98572	0.98592	0.98592
188	0.98500	0.9853	0.98538	0.98503	0.98533	0.98538
187	0.98440	0.9847	0.98469	0.98415	0.98445	0.98469
186	0.98360	0.9838	0.98418	0.98388	0.98418	0.98418
185	0.98310	0.9835	0.98351	0.98339	0.98344	0.98350
184	0.98230	0.9830	0.98300	0.9822	0.9827	0.98299
183	0.98190	0.9822	0.98226	0.98147	0.98221	0.98226
182	0.98110	0.9815	0.98152	0.97969	0.98152	0.98152
181	0.98020	0.9807	0.98103	0.97928	0.98103	0.98103
180	0.97970	0.9803	0.98029	0.97833	0.98029	0.98029
179	0.97910	0.9795	0.97951	0.97806	0.97951	0.97950
178	0.97830	0.9784	0.97840	0.97688	0.97821	0.97840
177	0.97720	0.9776	0.97747	0.9754	0.97724	0.97760
176	0.97640	0.9765	0.97669	0.97498	0.97669	0.97669
175	0.97530	0.9757	0.97571	0.9735	0.97571	0.97571
174	0.97435	0.9749	0.97479	0.97233	0.97469	0.97493
173	0.97362	0.9738	0.97383	0.97053	0.97376	0.97383
172	0.97266	0.9730	0.97303	0.96923	0.97303	0.97303
171	0.97186	0.9719	0.97193	0.9679	0.97193	0.97193
170	0.97076	0.9708	0.97076	0.96678	0.97076	0.97076
169	0.96922	0.9693	0.96929	0.96561	0.96929	0.96929
168	0.96813	0.9681	0.96813	0.96415	0.96813	0.96813
167	0.96634	0.9663	0.96634	0.96299	0.96634	0.96634
166	0.96504	0.9650	0.96504	0.96121	0.96504	0.96504
165	0.96371	0.9637	0.96371	0.95992	0.96371	0.96371
164	0.96242	0.9624	0.96242	0.9586	0.96242	0.96242
163	0.96064	0.9606	0.95998	0.95732	0.96064	0.96064
162	0.95912	0.9592	0.95821	0.95555	0.95919	0.95919
161	0.95803	0.9580	0.95692	0.9541	0.95804	0.95803
160	0.95567	0.9557	0.9556	0.95295	0.95571	0.95571
159	0.95432	0.9546	0.95433	0.9508	0.95457	0.95456

Source: Y. Nakagawa, and S. Miyazaki. 1981. Surrogate constraints algorithm for reliability optimization problems with two constraints. *IEEE Transactions on Reliability*, R-30, 175–180.

GA, genetic algorithm; ACO, ant colony optimization; and TS, tabu search.

prior knowledge nor exploit local search information. Compared to traditional meta-heuristic methods, a set of promising algorithms, hybrid GAs [33,34,48–50,132,133], are attractive since they retain the advantages of GAs in robustness and feasibility but significantly improve their computational efficiency and searching ability in finding global optimum with combining heuristic algorithms, neural network techniques, steepest decent methods or other local search methods.

For reliability optimization problems, exact solutions are not necessarily desirable because it is generally difficult to develop exact methods for reliability optimization problems which are equivalent to methods used for nonlinear integer programming problems [45]. However, exact methods may be particularly advantageous when the problem is not large. And more importantly, such methods can be used to measure the performance of heuristic or meta-heuristic methods.

10.6 Conclusions and Discussions

We have reviewed the recent research on optimal reliability design. Many publications have addressed this problem using different system structures, performance measures, problem formulations and optimization techniques.

The systems considered here mainly include series-parallel systems, k-out-of-n: G systems, bridge networks, n-version programming architecture, recovery block architecture and other unspecified coherent systems. The recently introduced NVP and RB belong to the category of fault tolerant architecture, which usually considers both software and hardware.

Reliability is still employed as a system performance measure in a majority of cases, but percentile life does provide a new perspective on optimal design without the requirement of a specified mission time. Availability is primarily used as the performance measure of renewable multi-state systems whose optimal design has been emphasized and well developed in the past 10 years. Optimal design problems are generally formulated to maximize an appropriate system performance measure under resource constraints, and more realistic problems involving multi-objective programming are also being considered.

When turning to optimization techniques, heuristic, meta-heuristic and exact methods are significantly applied in optimal reliability design. Recently, many advances in meta-heuristics and exact methods have been reported. Particularly, a new meta-heuristic method called ant colony optimization has been introduced and demonstrated to be a very promising general method in this field. Hybrid GAs may be the most important recent development among the optimization techniques since they retain the advantages of GAs in robustness and feasibility but significantly improve their computational efficiency.

Optimal reliability design has attracted many researchers, who have produced hundreds of publications since 1960. Due to the increasing complexity of practical engineering systems and the critical importance of reliability in these complex systems, this still seems to be a very fruitful area for future research.

Compared to traditional binary-state systems, there are still many unsolved topics in MSS optimal design including:

(a) using percentile life as a system performance measure
(b) involving cold-standby redundancy
(c) nonrenewable MSS optimal design
(d) applying optimization algorithms other than GAs, especially hybrid optimization techniques

From the view of optimization techniques, there are opportunities for improved effectiveness and efficiency of reported ACO, TS, IA and GDA, while some new meta-heuristic algorithms such as harmony search algorithm and particle swarm optimization may offer excellent solutions for reliability optimization problems. Hybrid optimization techniques are another very promising general developmental direction in this field. They may combine heuristic methods, NN or some local search methods with all kinds of meta-heuristics to improve computational efficiency or with exact methods to reduce search space. We may even be able to combine two meta-heuristic algorithms such as GA and SA or ACO.

The research dealing with the understanding and application of reliability at the nano level has also demonstrated its attraction and vitality in recent years. Optimal system design that considers reliability within the uniqueness of nano-systems has seldom been reported in the literature. It deserves a lot more attention in the future. In addition, uncertainty and component dependency will be critical areas to consider in future research on optimal reliability design.

References and Bibliography

1. F. Altiparmak, B. Dengiz, and A. E. Smith. 2003. Optimal design of reliable computer networks: a comparison of metaheuristics. *Journal of Heuristics*, 9, 471–487.
2. S. V. Amari, J. B. Dugan, and R. B. Misra. 1999. Optimal reliability of systems subject to imperfect fault-coverage. *IEEE Transactions on Reliability*, 48, 275–284.
3. S. V. Amari, H. Pham, and G. Dill. 2004. Optimal design of k-out-of-n: G subsystems subjected to imperfect fault-coverage. *IEEE Transactions on Reliability*, 53, 567–575.
4. J. D. Andrews, and L. A. Bartlett. 2005. A branching search approach to safety system design optimization. *Reliability Engineering and System Safety*, 87, 23–30.
5. Y. P. Aneja, R. Chandrasekaran, and K. P. K. Nair. 2004. Minimal-cost system reliability with discrete-choice sets for components. *IEEE Transactions on Reliability*, 53, 71–76.
6. V. C. Bueno. 2005. Minimal standy redundancy allocation in a k-out-of-n: F system of dependent components. *European Journal of Operation Research*, 165, 786–793.
7. P. G. Busacca, M. Marseguerra, and E. Zio. 2001. Multiobjective optimization by genetic algorithms: application to safety systems. *Reliability Engineering and System Safety*, 72, 59–74.
8. M. S. Chen. 1987. On the computational complexity of reliability redundancy allocation in a series system. *Operation Research Letter*, SE-13, 582–592.
9. T.-C. Chen, and P-S. You. 2005. Immune algorithms-based approach for redundant reliability problems with multiple component choices. *Computers in Industry*, 56, 195–205.
10. D. W. Coit, and A. E. Smith. 1996. Reliability optimization of series-parallel systems using a genetic algorithm. *IEEE Transactions on Reliability*, 45, 254–260.
11. D. W. Coit, and A. E. Smith. 1998. Redundancy allocation to maximize a lower percentile of the system time-to-failure distribution. *IEEE Transactions on Reliability*, 47, 79–87.
12. D. W. Coit, and J. Liu. 2000. System reliability optimization with *k*-out-of-*n* subsystems. *International Journal of Reliability, Quality and Safety Engineering*, 7, 129–142.
13. D. W. Coit. 2001. Cold-standby redundancy optimization for nonrepairable systems. *IIE Transactions,* 33, 471–478.
14. D. W. Coit, and A. E. Smith. 2002. Genetic algorithm to maximize a lower-bound for system time-to-failure with uncertain component Weibull parameters. *Computer and Industrial Engineering*, 41, 423–440.
15. D. W. Coit. 2003. Maximization of system reliability with a choice of redundancy strategies. *IIE Transactions*, 35, 535–543.
16. D. W. Coit, T. Jin, and N. Wattanapongsakorn. 2004. System optimization with component reliability estimation uncertainty: a multi-criteria approach. *IEEE Transactions on Reliability*, 53, 369–380.
17. L. Cui, W. Kuo, H. T. Loh, and M. Xie. 2004. Optimal allocation of minimal & perfect repairs under resource constraints. *IEEE Transactions on Reliability*, 53, 193–199.
18. C. K. Dimou, and V. K. Koumousis. 2003. Competitive genetic algorithms with application to reliability optimal design. *Advances in Engineering Software*, 34, 773–785.
19. M. Djerdjour. 1997. An enumerative algorithm framework for a class of nonlinear programming problems. *European Journal of Operation Research*, 101, 101–121.
20. M. Djerdjour, and K. Rekab. 2001. A branch and bound algorithm for designing reliable systems at a minimum cost. *Applied Mathematics and Computation*, 118, 247–259.
21. M. Dorigo. 1992. Optimization learning and natural algorithm. Ph.D. Dissertation, Politecnico di Milano, Italy.
22. M. Dorigo, V. Maniezzo, and A. Colorni. 1996. Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man and Cybernetics. Part B, Cybernetics,* 26, 29–41.
23. M. Dorigo, and L. M. Gambardella. 1997. Ant colonies for traveling salesman problem. *Biosystem*, 43, 73–81.

24. M. Dorigo, and L. M. Gambardella. 1997. Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Computers,* 1, 53–66.
25. M. Dorigo, and G. Di Caro. 1999. The ant colony optimization metaheuristic. In *New Ideas in Optimization*, D. Corne, M. Dorigo, and F. Glover, Eds. Maidenhead, UK: McGraw-Hill, 11–32.
26. M. Dorigo, G. Di Caro, and L. M. Gambardella. 1999. Ant algorithm for discrete optimization. *Artificial Life*, 5(2), 137–172.
27. C. Elegbede, C. Chu, K. Adjallah, and F. Yalaoui. 2003. Reliability allocation through cost minimization. *IEEE Transactions on Reliability*, 52, 106–111.
28. C. Elegbede, and K. Adjallah. 2003. Availability allocation to repairable systems with genetic algorithms: a multi-objective formulation. *Reliability Engineering and System Safety*, 82, 319–330.
29. M. Gen, and J. R. Kim. 1999. GA-based reliability design: state-of-the-art survey. *Computer and Industrial Engineering*, 37, 151–155.
30. C. Ha, and W. Kuo. 2006. Reliability redundancy allocation: an improved realization for nonconvex nonlinear programming problems. *European Journal of Operational Research*, 171, 24–38.
31. C. Ha, and W. Kuo. Multi-paths iterative heuristic for redundancy allocation: The tree heuristic. *IEEE Transactions on Reliability*, to be published.
32. C. Ha, and W. Kuo. 2005. Multi-path approach for reliability-redundancy allocation using a scaling method. *Journal of Heuristics*, 11, 201–217.
33. C. C. Hsieh, and Y. C. Hsieh. 2003. Reliability and cost optimization in distributed computing systems. *Computer and Operations Research*, 30, 1103–1119.
34. C. C. Hsieh. 2003. Optimal task allocation and hardware redundancy policies in distributed computing system. *European Journal of Operational Research*, 147, 430–447.
35. Y. C. Hsieh, T. C. Chen, and D. L. Bricker. 1998. Genetic algorithm for reliability design problems. *Microelectronics Reliability*, 38, 1599–1605.
36. Y. C. Hsieh. 2002. A linear approximation for redundant reliability problems with multiple component choices. *Computer and Industrial Engineering*, 44, 91–103.
37. J. Huang, M. J. Zuo, and Y. Wu. 2000. Generalized multi-state k-out-of-n: G systems. *IEEE Transactions on Reliability*, 49, 105–111.
38. T. Jin, and D. W. Coit. 2001. Variance of system-reliability estimates with arbitrarily repeated components. *IEEE Transactions on Reliability*, 50, 409–413.
39. K. O. Kim, and W. Kuo. 2003. Percentile life and reliability as a performance measures in optimal system design. *IIE Transactions*, 35, 1133–1142.
40. S. Kulturel-Kotz, C. D. Lai, and M. Xie. 2003. On the effect of redundancy for systems with dependent components. *IIE Transactions*, 35, 1103–1110.
41. S. Kulturel-Konak, A. E. Smith, and D. W. Coit. 2003. Efficiently solving the redundancy allocation problem using tabu search. *IIE Transactions*, 35, 515–526.
42. U. D. Kumar, and J. Knezevic. 1998. Availability based spare optimization using renewal process. *Reliability Engineering and Systems Safety*, 59, 217–223.
43. W. Kuo, C. L. Hwang, and F. A. Tillman. 1978. A note on heuristic method for in optimal system reliability. *IEEE Transactions on Reliability*, 27, 320–324.
44. W. Kuo, K. Chien, and T. Kim. 1998. *Reliability, Yield and Stress Burn-in*. Kluwer.
45. W. Kuo, and V. Prasad. 2000. An annotated overview of system-reliability optimization. *IEEE Transactions on Reliability*, 49, 487–493.
46. W. Kuo, V. R. Prasad, F. A. Tillman, and C. L. Hwang. 2001. *Optimal Reliability Design: Fundamentals and Applications*. New York: Cambridge University Press.
47. W. Kuo, and M. J. Zuo. 2003. *Optimal Reliability Modeling: Principles and Applications*. John Wiley & Sons.
48. C. Y. Lee, M. Gen, and W. Kuo. 2002. Reliability optimization design using hybridized genetic algorithm with a neural network technique. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E84-A, 627–637.

49. C. Y. Lee, Y. Yun, and M. Gen. 2002. Reliability optimization design using hybrid NN-GA with fuzzy logic controller. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E85-A, 432–447.
50. C. Y. Lee, M. Gen, and Y. Tsujimura. 2002. Reliability optimization design for coherent systems by hybrid GA with fuzzy logic controller and local search. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E85-A, 880–891.
51. H. Lee, W. Kuo, and C. Ha. 2003. Comparison of max-min approach and NN method for reliability optimization of series-parallel system. *Journal of System Science and Systems Engineering*, 12, 39–48.
52. G. Levitin, A. Lisnianski, H. Ben-Haim, and D. Elmakis. 1998. Redundancy optimization for series-parallel multi-state systems. *IEEE Transactions on Reliability*, 47(2), 165–172.
53. G. Levitin, and A. Lisnianski. 1998. Structure optimization of power system with bridge topology. *Electric Power Systems Research*, 45, 201–208.
54. G. Levitin, and A. Lisnianski. 1999. Joint redundancy and maintenance optimization for multistate series-parallel systems. *Reliability Engineering and System Safety*, 64, 33–42.
55. G. Levitin, and A. Lisnianski. 1999. Importance and sensitivity analysis of multi-state systems using the universal generating function method. *Reliability Engineering and System Safety*, 65, 271–282.
56. G. Levitin, and A. Lisnianski. 2000. Survivability maximization for vulnerable multi-state systems with bridge topology. *Reliability Engineering and System Safety*, 70, 125–140.
57. G. Levitin. 2001. Redundancy optimization for multi-state system with fixed resource-requirements and unreliable sources. *IEEE Transactions on Reliability*, 50, 52–59.
58. G. Levitin, and A. Lisnianski. 2001. Reliability optimization for weighted voting system. *Reliability Engineering and System Safety*, 71, 131–138.
59. G. Levitin, and A. Lisnianski. 2001. Structure optimization of multi-state system with two failure modes. *Reliability Engineering and System Safety*, 72, 75–89.
60. G. Levitin, and A. Lisnianski. 2001. Optimal separation of elements in vulnerable multi-state systems. *Reliability Engineering and System Safety*, 73, 55–66.
61. G. Levitin. 2001. Reliability evalution for acyclic conseucutively connected network with multistate elements. *Reliability Engineering and System Safety*, 73, 137–143.
62. G. Levitin, and A. Lisnianski. 2001. A new approach to solving problems of multi-state system reliability optimization. *Quality and Reliability Engineering International*, 17, 93–104.
63. G. Levitin. 2002. Optimal allocation of multi-state retransmitters in acyclic transmission networks. *Reliability Engineering and System Safety*, 75, 73–82.
64. G. Levitin. 2002. Optimal allocation of elements in a linear multi-state sliding window system. *Reliability Engineering and System Safety*, 76, 245–254.
65. G. Levitin. 2002. Optimal series-parallel topology of multi-state system with two failure modes. *Reliability Engineering and System Safety*, 77, 93–107.
66. G. Levitin. 2002. Evaluating correct classification probability for weighted voting classifiers with plurality voting. *European Journal of Operational Research*, 141, 596–607.
67. G. Levitin. 2003. Optimal allocation of multi-state elements in linear consecutively connected systems with vulnerable nodes. *European Journal of Operational Research*, 150, 406–419.
68. G. Levitin. 2003. Optimal allocation of multistate elements in a linear consecutively-connected system. *IEEE Transactions on Reliability,* 52, 192–199.
69. G. Levitin. 2003. Reliability evaluation for acyclic transmission networks of multi-state elements with delays. *IEEE Transactions on Reliability,* 52, 231–237.
70. G. Levitin. 2003. Linear multi-state sliding-window systems. *IEEE Transactions on Reliability,* 52(2), 263–269.
71. G. Levitin. 2003. Reliability of multi-state systems with two failure-modes. *IEEE Transactions on Reliability*, 52, 340–348.

72. G. Levitin, and A. Lisnianski. 2003. Optimizing survivability of vulnerable series-parallel multi-state systems. *Reliability Engineering and System Safety*, 79, 319–331.
73. G. Levitin. 2003. Optimal multilevel protection in series-parallel systems. *Reliability Engineering and System Safety*, 81, 93–102.
74. G. Levitin, Y. Dai, M. Xie, and K. L. Poh. 2003. Optimizing survivability of multi-state systems with multi-level protection by multi-processor genetic algorithm. *Reliability Engineering and System Safety*, 82, 93–104.
75. G. Levitin. 2004. A universal generating function approach for the analysis of multi-state systems with dependent elements. *Reliability Engineering and System Safety*, 84, 285–292.
76. G. Levitin. 2004. Consecutive k-out-of-r-from-n system with multiple failure criteria. *IEEE Transactions on Reliability*, 53, 394–400.
77. G. Levitin. 2004. Reliability optimization models for embedded systems with multiple applications. *IEEE Transactions on Reliability*, 53, 406–416.
78. G. Levitin. 2005. Uneven allocation of elements in linear multi-state sliding window system. *European Journal of Operational Research*, 163, 418–433.
79. G. Levitin. 2005. Optimal structure of fault-tolerant software systems. *Reliability Engineering and System Safety*, 89, 286–295.
80. G. Levitin. Reliability, and performance analysis of hardware-software systems with fault-tolerant software components. *Reliability Engineering and System Safety*, 91, 570–579.
81. Y. C. Liang, and A. E. Smith. 2004. An ant colony optimization algorithm for the redundancy allocation problem. *IEEE Transactions on Reliability*, 53, 417–423.
82. F. Lin, W. Kuo, and F. Hwang. 1999. Structure importance of consecutive-k-out-of-n systems. *Operations Research Letters*, 25, 101–107.
83. F. Lin, and W. Kuo. 2002. Reliability importance and invariant optimal allocation. *Journal of Heuristics*, 8, 155–172.
84. A. Lisnianski, G. Levitin, and H. Ben-Haim. 2000. Structure optimization of multi-state system with time redundancy. *Reliability Engineering and System Safety*, 67, 103–112.
85. A. Lisianski, and G. Levitin. 2003. *Multi-state System Reliability. Assessment, Optimization and Applications*. World Scientific.
86. P. X. Liu, M. J. Zuo, and M. Q-H. Meng. 2003. Using neural network function approximation for optimal design of continuous-state parallel-series systems. *Computers and Operations Research*, 30, 339–352.
87. M. R. Lyu, S. Rangarajan and A. P. A. van Moorsel. 2002. Optimal allocation of test resources for software reliability growth modeling in software development. *IEEE Transactions on Reliability*, 51, 183–192.
88. V. Maniezzo, L. M. Gambardella, and F. De Luigi. 2004. Ant colony optimization. In *New Optimization Techniques in Engineering*, G. C. Onwubolu, and B. V. Babu, Eds. Springer-Verlag Berlin Heidelberg, 101–117.
89. M. Marseguerra, and E. Zio. 2000. Optimal reliability/availability of uncertain systems via multi-objective genetic algorithms. *2000 Proceedings Annual Reliability and Maintainability Symposium*, 222–227.
90. M. Marseguerra, and E. Zio. 2004. System design optimization by genetic algorithm. *IEEE Transactions on Reliability*, 53, 424–434.
91. M. Marseguerra, E. Zio, L. Podofillini, and D. C. Coit. 2005. Optimal design of reliable network systems in presence of uncertainty. *IEEE Transactions on Reliability*, 54, 243–253.
92. Y. Massim, A. Zeblah, R. Meziane, M. Benguediab, and A. Ghouraf. 2005. Optimal design and reliability evaluation of multi-state series-parallel power systems. *Nonlinear Dynamics*, 40, 309–321.
93. K. B. Misra. 1986. On optimal reliability design: a review. *System Science*, 12, 5–30.
94. K. B. Misra. 1991. An algorithm to solve integer programming problems: an efficient tool for reliability design. *Microelectronics and Reliability*, 31, 285–294.

95. N. Nahas, and M. Nourelfath. 2005. Ant system for reliability optimization of a series system with multiple-choice and budget constraints. *Reliability Engineering and System Safety*, 87, 1–12.
96. Y. Nakagawa, and S. Miyazaki. 1981. Surrogate constraints algorithm for reliability optimization problems with two constraints. *IEEE Transactions on Reliability*, R-30, 175–180.
97. K. Y. K. Ng, and N. G. F. Sancho. 2001. A hybrid dynamic programming/depth-first search algorithm with an application to redundancy allocation. *IIE Transactions*, 33, 1047–1058.
98. L. Nordmann, and H. Pham. 1999. Weighted voting systems. *IEEE Transactions on Reliability*, 48, 42–49.
99. M. Nourelfath, and Y. Dutuit. 2004. A combined approach to solve the redundancy optimization problem for multi-state systems under repair policies. *Reliability Engineering and System Safety*, 84, 205–213.
100. V. R. Prasad, and M. Raghavachari. 1998. Optimal allocation of interchangeable components in a series-parallel system. *IEEE Transactions on Reliability*, 47, 255–260.
101. V. R. Prasad, W. Kuo, and K. O. Kim. 2000. Optimal allocation of s-identical multi-functional spares in a series system. *IEEE Transactions on Reliability*, 49, 323–330.
102. V. R. Prasad, and W. Kuo. 2000 Reliability optimization of coherent systems. *IEEE Transactions on Reliability*, 49, 323–330.
103. V. R. Prasad, W. Kuo, and K. O. Kim. 2001. Maximization of a percentile life of a series system through component redundancy allocation. *IIE Transactions,* 33, 1071–1079.
104. J. E. Ramirez-Marquez, D. W. Coit, and A. Konak. 2004. Redundancy allocation for series-parallel systems using a max-min approach. *IIE Transactions*, 36, 891–898.
105. J. E. Ramirez-Marquez, and D. W. Coit. 2004. A heuristic for solving the redundancy allocation problem for multi-state series-parallel system. *Reliability Engineering and System Safety*, 83, 341–349.
106. V. Ravi, P. J. Reddy, and H. J. Zimmermann. 2000. Fuzzy global optimization of complex system reliability. *IEEE Transactions on Fuzzy Systems*, 8, 241–248.
107. V. Ravi. 2004. Optimization of complex system reliability by a modified great deluge algorithm. *Asia-Pacific Journal of Operational Research*, 21, 487–497.
108. Y. Ren, and J. B. Dugan. 1998. Design of reliable system using static & dynamic fault trees. *IEEE Transactions on Reliability*, 47, 234–244.
109. C. M. Rocco, J. A. Moreno, and N. Carrasquero A cellular evolution approach applied to reliability optimization of complex systems. In *2000 Proceedings Annual Reliability and Maintainability Symposium*, Los Angeles, CA, 210–215.
110. R. Romera, J. E. Valdés, and R. I. Zequeira. 2004. Active-redundancy allocation in systems. *IEEE Transactions on Reliability*, 53, 313–318.
111. J. O. Royset, and E. Polak. 2004. Reliability-based optimal design using sample average approximations. *Probabilistic Engineering Mechanics*, 19, 331–343.
112. J. O. Royset, A. D. Kiureghian, and E. Polak. 2001. Reliability-based optimal design of series structural systems. *Journal of Engineering Mechanics*, June, 607–614.
113. H. S. Ryoo. 2005. Robust meta-heuristic algorithm for redundancy optimization in large-scale complex systems. *Annals of Operations Research*, 133, 209–228.
114. M. Sasaki, and M. Gen. 2003. A method of fuzzy multi-objective nonlinear programming with GUB structure by hybrid genetic algorithm. *International Journal of Smart Engineering Design*, 5, 281–288.
115. M. Sasaki, and M. Gen. 2003. Fuzzy multiple objective optimal system design by hybrid genetic algorithm. *Applied Soft Computing*, 5, 281–288.
116. P. S. Shelokar, V. K. Jayaraman, and B. D. Kulkarni. 2002. Ant algorithm for single and multiobjective reliability optimization problems. *Quality and Reliability Engineering International*, 18, 497–514.
117. R. Stocki, K. Kolanek, S. Jendo, and M. Kleiber. 2001. Study on discrete optimization techniques in reliability-based optimization of truss structures. *Computers and Structures*, 79, 2235–2247.

118. B. Suman. 2003. Simulated annealing-based multiobjective algorithm and their application for system reliability. *Engineering Optimization*, 35, 391–416.
119. C. S. Sung, and Y. K. Cho. 1999. Branch-and-bound redundancy optimization for a series system with multiple-choice constraints. *IEEE Transactions on Reliability*, 48, 108–117.
120. C. S. Sung, and Y. K. Cho. 2000. Reliability optimization of a series system with multiple-choice and budget constraints. *European Journal of Operational Research*, 127, 158–171.
121. T. Taguchi, T. Yokota, and M. Gen. 1998. Reliability optimal design problem with interval coefficients using hybrid genetic algorithms. *Computers and Industrial Engineering*, 35, 373–376.
122. T. Taguchi, and T. Yokota. 1999. Optimal design problem of system reliability with interval coefficient using improved genetic algorithms. *Computers and Industrial Engineering*, 37, 145–149.
123. F. A. Tillman, C. L. Hwang, and W. Kuo. 1977. Optimization techniques for system reliability with redundancy–a review. *IEEE Transactions on Reliability*, R-26, 148–152.
124. N. Wattanapongsakorn and S. P. Levitan. 2004. Reliability optimization models for embedded systems with multiple applications. *IEEE Transactions on Reliability*, 53, 406–416.
125. S. Wolfram. 1984. Cellular automata as models of complexity. *Nature*, 311, 419–424.
126. A. Yalaoui, C. Chu, and E. Châtelet. 2005. Reliability allocation problem in a series-parallel system. *Reliability Engineering and System Safety*, 90, 55–61.
127. A. Yalaoui, E. Châtelet, and C. Chu. 2005. A new dynamic programming method for reliability and redundancy allocation in a parallel-series system. *IEEE Transactions on Reliability*, 54, 254–261.
128. J. E. Yang, M. J. Hwang, T. Y. Sung, and Y. Jin. 1999. Application of genetic algorithm for reliability allocation in nuclear power plants. *Reliability Engineering and System Safety*, 65, 229–238.
129. P. S. You, and T. C. Chen. 2005. An efficient heuristic for series-parallel redundant reliability problems. *Computer & Operations Research*, 32, 2117–2127.
130. W. Y. Yun, and J. W. Kim. 2004. Multi-level redundancy optimization in series systems. *Computer and Industrial Engineering*, 46, 337–346.
131. E. P. Zafiropoulos, and E. N. Dialynas. 2004. Reliability and cost optimization of electronic devices considering the component failure rate uncertainty. *Reliability Engineering and System Safety*, 84, 271–284.
132. R. Zhao, and B. Liu. 2003. Stochastic programming models for general redundancy-optimization problems. *IEEE Transactions on Reliability*, 52, 181–191.
133. R. Zhao, and B. Liu. 2004. Redundancy optimization problems with uncertainty of combining randomness and fuzziness. *European Journal of Operation Research*, 157, 716–735.

11

Lower Confidence Bounds for System Reliability from Binary Failure Data Using Bootstrapping

Lawrence M. Leemis
The College of William and Mary

11.1 Introduction

We consider the problem of determining a $(1-\alpha)$ 100% lower confidence bound on the system reliability for a coherent system of k components using the failure data (y_i, n_i), where y_i is the number of components of type i that pass the test and n_i is the number of components of type i on test, $i = 1, 2, \ldots, k$. We assume throughout that the components fail independently, e.g. no common-cause failures. The outline of the article is as follows. We begin with the case of a single ($k = 1$) component system where n components are placed on a test and y components pass the test. The Clopper–Pearson lower bound is used to provide a lower bound on the reliability. This model is then generalized to the case of multiple ($k > 1$) components. Bootstrapping is used to estimate the lower confidence bound on system reliability. We then address a weakness in the bootstrapping approach—the fact that the sample size is moot in the case of perfect test results, e.g. when $y_i = n_i$ for some i. This weakness is overcome by using a beta prior distribution to model the component reliability before performing the bootstrapping. Two subsections consider methods for estimating the parameters in the beta prior distribution for components with perfect test results. The first subsection considers the case when previous test results are available, and the second subsection considers the case when no previous test results are available. A simulation study compares various algorithms for calculating a lower confidence bound on the system reliability. The last section contains conclusions.

11.2 Single-component Systems

Single-component systems are considered first because (1) there are known approximate confidence intervals for the lower reliability confidence bound; and (2) these intervals will be used later in the chapter to help determine the appropriate parameters for the beta distribution in the case where no prior test results exist on the component of interest.

Let n components be placed on test and let y of these components pass the test. Under the assumption that the test values (1 for pass, 0 for failure) $X_1, X_2, \ldots, X_n$ are independent and identically distributed Bernoulli random variables with unknown parameter p, $Y = \sum_{i=1}^{n} X_i$ is a binomial random variable with parameters n and p. The maximum likelihood estimator for p is $\hat{p} = Y/n$, which is unbiased and consistent. The interest here is in a lower confidence bound for the reliability p.

There is a wide literature on confidence intervals of this type since a confidence interval on a proportion is of interest on anything from a political poll to consumer preference. Vollset (1993) compares 13 confidence intervals and Newcombe (2001) compares seven confidence intervals. Rather than fine-tuning these intervals as has been suggested by many authors, we have settled on using the Clopper–Pearson (CP) "exact" interval even though Newcombe (2001) points out that its status as a gold standard has been disputed recently because the method is conservative, i.e. the actual coverage is greater than or equal to the stated coverage (see Agresti and Coull, 1998, for details).

Let $p_L < p < p_U$ be an "exact" (Blyth 1986) CP two-sided confidence interval for p, where p_L and p_U are functions of the sample size n, the number of successes y, and the stated coverage of the interval, $1-\alpha$. This is an approximate confidence interval due to the discrete nature of the binomial distribution. For $y = 1, 2, \ldots, n-1$ the lower limit p_L satisfies (see, for example, Agresti and Coull 1998).

$$\sum_{k=y}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \alpha/2.$$

For $y = 1, 2, \ldots, n-1$, the upper limit p_U satisfies

$$\sum_{k=0}^{y} \binom{n}{k} p_U^k (1-p_U)^{n-k} = \alpha/2.$$

As shown in Leemis and Trivedi (1996), these confidence interval limits can be expressed in terms of quantiles of the F distribution:

$$\frac{1}{1+\dfrac{n-y+1}{yF_{2y,2(n-y+1),1-\alpha/2}}} < p < \frac{1}{1+\dfrac{n-y}{(y+1)F_{2(y+1),2(n-y),\alpha/2}}},$$

where the third subscript on F refers to the right-hand tail probability.

Simply reallocating the probability α to the lower limit gives the following lower confidence bound for the reliability:

$$p_L = \frac{1}{1+\frac{n-y+1}{yF_{2y,2(n-y+1),1-\alpha}}},$$

for $y = 1, 2, \ldots, n-1$. For the case of all failures ($y = 0$), the lower bound is, of course, $p_L = 0$. For the case of all passes ($y = n$), the lower bound is $p_L = \alpha^{1/n}$.

Example: CP Lower Confidence Interval Bounds

The following four sets of values for n and y give point estimates and 95% CP lower confidence interval bounds for the reliability:

$$n = 10, y = 7 \quad \Rightarrow \quad \hat{p} = 0.7, p_L = 0.393.$$

$$n = 100, y = 97 \quad \Rightarrow \quad \hat{p} = 0.97, p_L = 0.924.$$

$$n = 10, y = 10 \quad \Rightarrow \quad \hat{p} = 1.0, p_L = 0.741.$$

$$n = 100, y = 100 \quad \Rightarrow \quad \hat{p} = 1.0, p_L = 0.970.$$

An S-Plus function named confintlower is given in Appendix 11.1 which can be used to calculate these lower confidence interval bounds. Figure 11.1 is a plot of y vs. p_L when $n = 10$ for $\alpha = 0.10, 0.05, 0.01$, with the points connected with line segments. Figure 11.2 contains a similar plot for $n = 100$. The lower bounds are monotonic in y, n, and α.

11.3 Multiple-component Systems

A three-component ($k = 3$) series system is used as an example throughout this section, although the techniques described here apply to any coherent system of k independent components. The number of components tested and the number of passes for each type of component for the example are given in Table 11.1. The point estimate for the system reliability is

$$\frac{21}{23} \cdot \frac{27}{28} \cdot \frac{82}{84} = \frac{1107}{1288} \cong 0.8595.$$

The remainder of this section involves the use of bootstrapping (Efron and Tibshirani 1993) to calculate a lower 95% confidence interval bound. Other authors (e.g. Martin 1990 and Padgett and Tomlinson 2003) have used bootstrapping for determining confidence limits. The approach used here differs conceptually from the standard bootstrap problem, where the standard error of a *single* unknown distribution is estimated by resampling iid data. In our setting, there are *k different* distributions (one for each component) and we resample component reliabilities and combine using the reliability function to yield the system reliability.

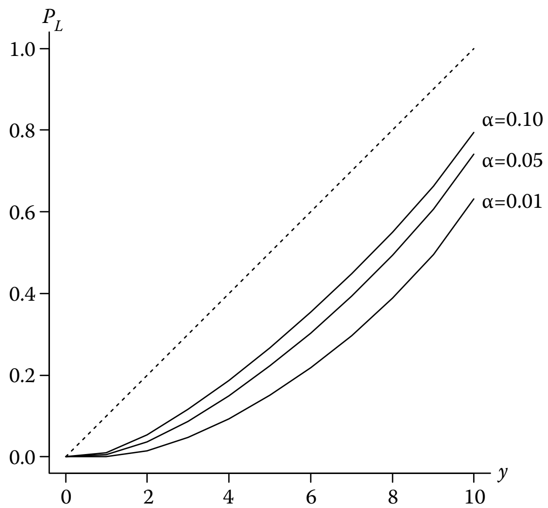


FIGURE 11.1 Point estimate (dashed) and CP lower confidence bounds (solid) when $n = 10$.

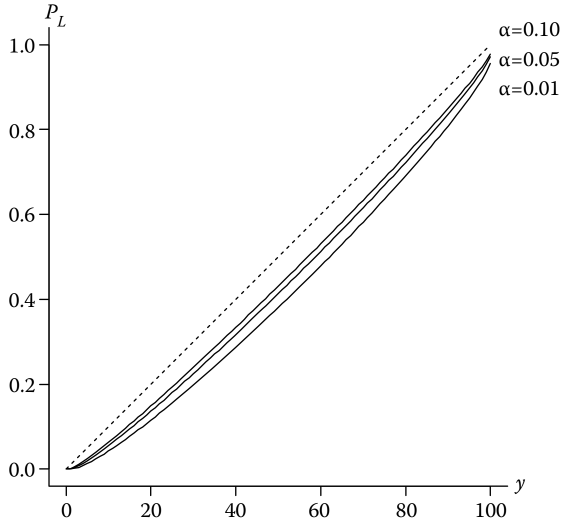


FIGURE 11.2 Point estimate (dashed) and CP lower confidence bounds (solid) when $n=100$.

TABLE 11.1 Failure Data for a Three-component Series System Example

Component number	$i=1$	$i=2$	$i=3$
Number passing (y_i)	21	27	82
Number on test (n_i)	23	28	84

Bootstrapping resamples B of the systems, calculates the system reliability, then outputs the αB^{th} ordered system reliability. More specifically, for the three-component system of interest, the bootstrapping algorithm follows the following steps.

- For the first component, the data set (21 ones and two zeros) is sampled with replacement 23 times.
- These values are summed and divided by 23 yielding a reliability estimate for the first component.
- The previous two steps are repeated for components 2 and 3.
- The product of the reliability estimates for the three components are multiplied (because the components are arranged in series and their failures are independent) to give a system reliability estimate.

The above procedure is repeated B times. The B system reliability estimates are then sorted. Finally, the αB^{th} ordered system reliability is output, which is used as a lower bound on the system reliability.

The algorithm for estimating the (1–α)100% lower confidence interval bound is given in Table 11.2, where $\tilde{p}_i$ is a bootstrap estimate for the reliability of component i and z_j is a bootstrap estimate of the system reliability. The binomial distribution is appropriate since the resampling from the data set is performed with replacement. In the pseudocode in Table 11.2, indentation is used to indicate begin-end blocks. The returned value $z_{\alpha B}$ is the order statistic associated with the z_j's generated in the outside loop. See Law and Kelton (2000) for handling the case when αB is not an integer.

This algorithm has been implemented in S-Plus as a function named `seriessystemboot` which is given in Appendix 11.2. The first two arguments, `n` and `y`, are vectors of length k, and the third argument, `alpha`, is a real number between 0 and 1, e.g.

```
seriessystemboot(c (23, 28, 84), c(21, 27, 82), 0.05)
```

TABLE 11.2 Bootstrap Algorithm for Calculating a (1−α)100% Lower Confidence Bound for the Reliability of a k-component Series System

for j from 1 to B	[resampling loop]
for i from 1 to k	[loop through components]
$\tilde{p}_i \leftarrow$ Binomial(n_i, y_i/n_i)/n_i	[component i reliability]
$z_j \leftarrow \prod_{i=1}^{k} \tilde{p}_i$	[calculate system reliability]
sort **z**	[sort the system reliability values]
return $z_{\alpha B}$	[return the estimate for the lower bound]

prints a point estimate and a 95% lower confidence interval bound on the system reliability for the three-component series system considered in this section. After a call to `set.seed(3)` to set the random number seed, five calls to `seriessystemboot` yield the following estimates for p_L:

$$0.7426057 \quad 0.7482993 \quad 0.7456744 \quad 0.7486690 \quad 0.7453416.$$

The dispersion associated with these five estimates is due to the finite choice of B, i.e. $B=10{,}000$.

Resampling error can be eliminated using the symbolic Maple-based APPL (Glen et al. 2001). The APPL statements given in Appendix 11.3 utilize the `Product` and `Transform` procedures. This alternative approach to determining a lower 95% bootstrap confidence interval bound for the system reliability is equivalent to using an infinite value for B. Since $\tilde{p}_i$ can assume any one of n_i+1 values, there are a possible $24 \cdot 29 \cdot 85 = 59{,}160$ potential mass values for the random variable `T` determined by the `Product` procedure. Of these, only 6633 are distinct since the `Product` procedure combines repeated values. Since the random variable `T` from the APPL code plays an analogous role to the vector **z** from the bootstrap algorithm in Table 11.2, the lower 95% bootstrap confidence interval bound is the 0.05 fractile of the distribution of `T`, which is $p_L = 6723/9016 \cong 0.746$. This result using APPL is consistent with the standard resampling approach for finding the lower confidence interval limit based on the five results presented earlier (one equals 6723/9016 exactly, two fall above 6723/9016, and two fall below 6723/9016).

11.3.1 How Well Does the Bootstrap Procedure Perform?

This question is difficult to address since there is no exact interval to compare with even in the case of a single component. It is instructive, however, to isolate one component and compare the CP approach described in the previous section with bootstrapping. Arbitrarily choosing the second component with $n_2=28$ items on test, Figure 11.3 shows the CP lower confidence bound and the bootstrap lower confidence bound for $\alpha=0.05$ and for $y_2=10, 11, \ldots, 28$. The bootstrap lower confidence interval limit does not require any iteration, since the value plotted is $q/28$, where q is the smallest integer that satisfies

$$\sum_{k=0}^{q} \binom{n_2}{k} \left(\frac{y_2}{28}\right)^k \left(1-\left(\frac{y_2}{28}\right)\right)^{n_2-k} \geq \alpha,$$

i.e. q is the α-quantile of a binomial distribution with $n_2=28$ trials and probability of success $y_2/28$. Figure 11.3 shows that

- The bootstrap interval is more susceptible to the discrete nature of the binomial sampling scheme.
- The CP interval is wider than the bootstrap interval.

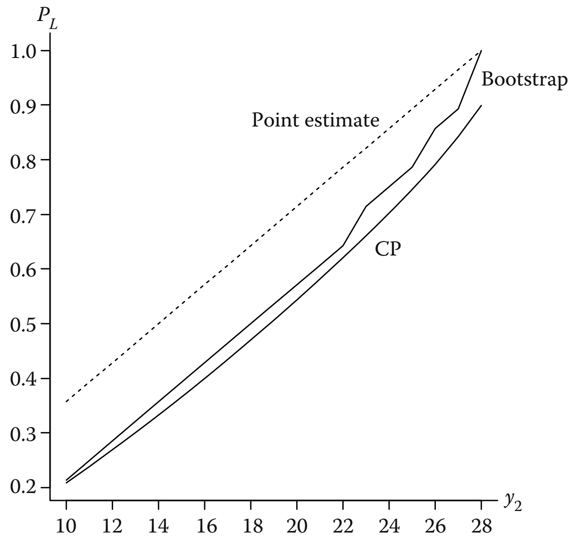


FIGURE 11.3 Point estimate (dashed), CP lower 95% confidence bound, and bootstrap lower 95% confidence bound for the reliability of component 2 based on a sample of size $n_2=28$.

Figure 11.3 also points out a glaring deficiency in the bootstrapping approach that was not revealed in the example in this section since all three of the system components had one or more failures during their life test. When component i has perfect test results (e.g. $y_i=n_i$), the sample size becomes irrelevant. Thus a test where two components out of two pass the test is equivalent to one where 100 components out of 100 pass the test from the perspective of the bootstrapping algorithm. This is clearly unacceptable. The next section gives a modification to the bootstrapping approach that adjusts for these perfect tests.

11.4 Perfect Component Test Results

The problem created by perfect component test results is likely to occur for components and systems with moderate to high reliability. As suggested by Chick (2001) and Martz and Waller (1982), a beta(α_1, α_2) prior distribution can be placed on the component reliability. The beta distribution is a logical choice for a prior distribution of the component reliability due to (1) the flexibility in the shape of its probability density function; (2) its (0, 1) support; and (3) its analytically tractable conjugate posterior distribution. Determining the values of the parameters α_1 and α_2 is a problem that will be addressed in the following two subsections.

The beta distribution has probability density function

$$f(x)=\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}x^{\alpha_1-1}(1-x)^{\alpha_2-1}, \qquad 0<x<1,$$

where α_1 and α_2 are positive shape parameters. This is the standard parameterization, although Martz and Waller (1982) use a slightly different form. The mean of a beta(α_1, α_2) random variable is

$$\mu=E[X]=\frac{\alpha_1}{\alpha_1+\alpha_2}$$

and the variance is

$$\sigma^2=V[X]=\frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}.$$

If the prior distribution of a reliability $P \sim \text{beta}(\alpha_1, \alpha_2)$ and the sampling is binomial (as it is in our case), then the posterior distribution of P is $\text{beta}(\alpha_1 + y, \alpha_2 + n - y)$, where n is the number of components on test and y is the number of passes. The difficulty in our case is in determining the appropriate values for α_1 and α_2. For the time being, we will proceed as if we know the values of α_1 and α_2 and give an algorithm for finding the lower reliability confidence bound p_L. Estimating α_1 and α_2 will be addressed subsequently.

A lower reliability confidence bound p_L can be determined by generating a bootstrap beta random variate (rather than a binomial) when the component test results are perfect. An algorithm for determining p_L for a k-component series system using B resamplings with some or all components having perfect tests is given Table 11.3. If component i has perfect test results (e.g. $y_i = n_i$) then the analyst must define the prior beta parameters α_{1i} and α_{2i}.

Example: Three-component Series System

Table 11.4 is identical to Table 11.1, except that component 2 now has perfect (28/28) test results. The point estimate for the system reliability increases to

$$\frac{21}{23} \cdot \frac{28}{28} \cdot \frac{82}{84} = \frac{41}{46} \cong 0.8913.$$

Thus the effect of the one additional component that passed the test increases the system reliability estimate from approximately 0.86 to approximately 0.89. This increase should be reflected in an appropriate increase in the lower confidence limit p_L.

This algorithm has been implemented as the S-Plus function `seriessystembayesboot` given in Appendix 11.4 (the number of bootstrap replications $B = 10{,}000$ and values of the beta parameters $\alpha_{12} = 1$ and $\alpha_{22} = 1$ are arbitrary). As before, n and y are vectors of length k, and `alpha` is a real number between 0 and 1, e.g.

```
seriessystembayesboot(c(23, 28, 84), c(21, 28, 82), 0.05)
```

prints a point estimate and a 95% lower confidence interval bound on the system reliability for the three-component series system. After a call to `set.seed(3)` to set the random number seed, five calls to `seriessystembayesboot` yield the following values for p_L:

```
0.7474437  0.7484738  0.7492014  0.7484301  0.7495972.
```

TABLE 11.3 Bootstrap Algorithm for Calculating a $(1 - \alpha)100\%$ Lower Confidence Bound for the Reliability of a k-component Series System when some Components have Perfect Test Results

for j from 1 to B	[resampling loop]
for i from 1 to k	[loop through components]
if $(y_i = n_i)$ $\tilde{p}_i \leftarrow \text{Beta}(\alpha_{1i} + y_i, \alpha_{2i})$	[component i reliability: perfect test]
else $\tilde{p}_i \leftarrow \text{Binomial}(n_i, y_i / n_i) / n_i$	[component i reliability: failure(s) occur]
$z_j \leftarrow \prod_{i=1}^{k} \tilde{p}_i$	[calculate system reliability]
sort $\mathbf{z}$	[sort the system reliability values]
return $z_{\alpha B}$	[return the estimate for the lower bound]

TABLE 11.4 Failure Data for a Three-component Series System

Component Number	$i = 1$	$i = 2$	$i = 3$
Number passing (y_i)	21	28	82
Number of test (n_i)	23	28	84

With the choice $\alpha_1 = \alpha_2 = 1$, the increase of approximately 0.03 in the point estimate of the system reliability from the previous example results in only a tiny increase in the lower confidence interval limits. This is clearly unacceptable. What happened? The arbitrary choice of $\alpha_1 = 1$ and $\alpha_2 = 1$ has resulted in a uniform prior distribution, which is an overly pessimistic assessment of the reliability of component 2, particularly in light of the perfect test results.

What choice would make more sense? It is important to skew the probability density function of the beta prior distribution so that its mean is greater than 1/2, or, equivalently, choose $\alpha_2 < \alpha_1$. There are four different shapes of the probability density function associated with the choice of parameters that satisfy the constraint $\alpha_2 < \alpha_1$. Most important is the value of the probability density function near $f(1)$, since these are the particular reliability values of interest. The following four cases demark various features of the probability density function.

- $f(0) = 0$ and $f(1) = 0$ when $1 < \alpha_2 < \alpha_1$ (Case I).
- $f(1)$ is finite when $1 = \alpha_2 < \alpha_1$ (Case II).
- a vertical asymptote at $x = 1$ and $f(0) > 0$ when $\alpha_2 < 1 = \alpha_1$ (Case III).
- a vertical asymptote at $x = 1$ and $f(0) = 0$ when $\alpha_2 < 1 < \alpha_1$ (Case IV).

We have disregarded the case $\alpha_2 < \alpha_1 < 1$ since this results in a vertical asymptote at both 0 and 1, which is inconsistent with the probability density function of a high reliability component. The most intuitively appealing of the four cases listed above it the fourth case, $\alpha_2 < 1 < \alpha_1$, since this minimizes the probability of generating a small beta variate (since $f(0) = 0$) and pushes as much of the probability near 1 as possible due to the vertical asymptote near 1.

Table 11.5 gives means of the beta prior distribution and lower confidence interval bounds for several combinations of α_1 and α_2 satisfying the constraint $\alpha_2 < \alpha_1$. The lower bounds are determined by taking the sample median of five runs of `seriessystembayesboot` with $B = 10{,}000$ resampled series systems per run. The subscript on the lower bound indicates which of the shapes in the list given above is represented. The value of the lower bound is quite sensitive to the choices of α_1 and α_2. There are many (α_1, α_2) pairs that yield a reasonable lower bound.

The following two subsections outline methods for estimating the parameters of the prior distribution. The first subsection considers the case when previous test results exist, so data are available to estimate $\hat{\alpha}_1$ and $\hat{\alpha}_2$. The second subsection considers the case when no previous test data are available.

11.4.1 Previous Test Data Exists

When previous test data that is representative of the current test data for a perfect component exists, this data can be fit to yield parameter estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ for the prior beta distribution. Let $z_1, z_2, \ldots, z_n$ denote the fraction surviving for previous tests on a component of interest with equal sample sizes (which has perfect test results and need a beta prior distribution). The maximum likelihood estimators satisfy the simultaneous equations (Evans et al. 2000):

$$\psi(\hat{\alpha}_1) - \psi(\hat{\alpha}_1 + \hat{\alpha}_2) = \frac{1}{n}\sum_{i=1}^{n} \log z_i,$$

TABLE 11.5 Prior Beta Distribution Mean and Lower 95% Confidence Interval Limit Estimate for the System Reliability

	α_2		
α_1	0.1	1	10
1	$0.909/0.779_{III}$	–	–
10	$0.990/0.783_{IV}$	$0.909/0.759_{II}$	–
100	$0.999/0.783_{IV}$	$0.990/0.779_{II}$	$0.909/0.725_{I}$

$$\psi(\hat{\alpha}_2) - \psi(\hat{\alpha}_1 + \hat{\alpha}_2) = \frac{1}{n}\sum_{i=1}^{n} \log(1 - z_i),$$

where ψ is the digamma function. Law and Kelton (2000) outline methods for calculating $\hat{\alpha}_1$ and $\hat{\alpha}_2$. These equations have no closed-form solution, and must be solved iteratively. Alternatively, the method of moments estimates are found by equating the population mean μ and population variance σ^2 to the associated sample moments:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i, \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n} (z_i - \bar{z})^2,$$

which results in the closed-form method of moments estimators:

$$\hat{\alpha}_1 = \frac{(1 - \bar{z})\bar{z}^2}{s^2} - \bar{z}, \qquad \hat{\alpha}_2 = \frac{\hat{\alpha}_1(1 - \bar{z})}{\bar{z}}.$$

Example: Estimating the Beta Parameters from Previous Experiments

Consider the previous example, where previous test results on component 2 have yielded the following $n = 4$ fractions surviving:

$$z_1 = \frac{27}{28}, \qquad z_2 = \frac{28}{28}, \qquad z_3 = \frac{26}{28}, \qquad z_4 = \frac{27}{28}.$$

Since the sample mean and variance are

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{27}{28} \cong 0.964, \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n} (z_i - \bar{z})^2 = \frac{1}{1568} \cong 0.000638,$$

the method of moments estimators are (these correspond to Case I from the previous list):

$$\hat{\alpha}_1 = \frac{1431}{28} \cong 51.11, \qquad \hat{\alpha}_2 = \frac{53}{28} \cong 1.89.$$

When these values for the parameters are used in `seriessystembayesboot`, the median of five lower 95% confidence bounds with $B = 10{,}000$ for the system reliability is 0.763.

11.4.2 No Previous Test Data Exists

We now turn to the more difficult case of determining the prior beta distribution parameter estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ in the case of a component with perfect test results and when no previous test data are available. For such a component, the point estimate of the component reliability is $\hat{p} = 1$ and the CP lower reliability bound is $p_L = \alpha^{1/n}$. One heuristic technique for determining the parameters is to choose $\hat{\alpha}_1$ and $\hat{\alpha}_2$ such that $F(p_L) = \alpha$, i.e.

$$\int_0^{p_L} \frac{\Gamma(\hat{\alpha}_1 + \hat{\alpha}_2)}{\Gamma(\hat{\alpha}_1)\Gamma(\hat{\alpha}_2)} x^{\hat{\alpha}_1 - 1}(1 - x)^{\hat{\alpha}_2 - 1} \mathrm{d}x = \alpha. \tag{11.1}$$

The intuition behind this choice is that 100α% of the time, a prior component reliability (which will be modified subsequently by the data set) will assume a value less than p_L. One problem with this criteria is that there are an infinite number of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ that satisfy this equation. Further refinement is necessary.

For a sample of size $n > 1$, the $(\hat{\alpha}_1, \hat{\alpha}_2)$ pair satisfying Equation 11.1 will (a) intersect the line $\hat{\alpha}_1 = 1$ on $0 < \hat{\alpha}_2 < 1$ and (b) intersect the line $\hat{\alpha}_2 = 1$ on $\hat{\alpha}_1 > 1$. One technique for determining a $(\hat{\alpha}_1, \hat{\alpha}_2)$ pair is to find the intersection of the values of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ that satisfy Equation 11.1 and the lines $\hat{\alpha}_1 = 1$ and $\hat{\alpha}_2 = 1$. These two points of intersection, or any point on the line segment connecting them can be used as prior beta distribution parameter estimates. It is interesting to note that

- The intersection of Equation 11.1 and the line $\hat{\alpha}_1 = 1$ corresponds to Case III for the beta distribution parameters (Scenario 1).
- The intersection of Equation 11.1 and the line $\hat{\alpha}_2 = 1$ corresponds to Case II for the beta distribution parameters (Scenario 2).
- Any point on the line segment connecting the two intersection points (not including the endpoints of the segment) corresponds to Case IV for the beta distribution parameters (Scenario 3).

We first consider the intersection of Equation 11.1 and $\alpha_1 = 1$. Integration of the beta probability density function is analytic in this case, yielding

$$1-(1-p_L)^{\hat{\alpha}_2} = \alpha$$

or

$$\hat{\alpha}_2 = \frac{\log(1-\alpha)}{\log(1-\alpha^{1/n})}.$$

Next, we first consider the intersection of Equation 11.1 and $\alpha_2 = 1$. The integration of the beta probability density function is analytic in this case as well, yielding

$$p_L^{\hat{\alpha}_1} = \alpha$$

or

$$\hat{\alpha}_1 = n.$$

Example: Three-component Series System with Beta Prior Distributions

Consider again the three-component series systems. System 1 has test results displayed in Table 11.1. System 2 has test results displayed in Table 11.4. The point estimate for the system reliability of System 1 is

$$\frac{21}{23}\cdot\frac{27}{28}\cdot\frac{82}{84} = \frac{1107}{1288} \cong 0.8595$$

and the point estimate for the system reliability of System 2 is

$$\frac{21}{23}\cdot\frac{28}{28}\cdot\frac{82}{84} = \frac{41}{46} \cong 0.8913.$$

Hence the slight difference between the two test results (the perfect test results for component 2) has resulted in a 0.8913–0.8595 = 0.0318 increase in the point estimate for the system reliability. A similar increase in the lower bound for the system reliability for a reasonable procedure is expected.

The earlier analysis of System 1 using APPL has resulted in an exact (no resampling variability) bootstrap 95% lower limit on the system reliability of 0.746. Table 11.6 contains 95% lower confidence limits for the system reliability using four different combinations of prior beta distribution parameter estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ for component 2. The parameter estimates for Scenario 3 are found by averaging the parameter estimates for Scenarios 1 and 2. The lower bounds p_L are determined by taking the median result of five runs with $B = 10{,}000$ replications using the bootstrap procedure described earlier. The column labeled Δp_L gives the difference between the lower confidence limit for System 2 and the lower confidence limit for System 1. The uniform prior model is too wide because the 0.748–0.746 = 0.002 increase in the lower bound is inconsistent with the 0.0318 increase in the point estimate for the system reliability. Based on this example only, Scenarios 1 and 3 seem to be the most appropriate since their increases in the lower bound bracket the increase in the point estimator for the system reliability.

TABLE 11.6 Lower Reliability Bounds (α=0.05) for the System Reliability of a Three-component Series System with Alternative Beta Prior Parameters

Model	$\hat{\alpha}_1$	$\hat{\alpha}_2$	p_L	Δp_L
Uniform prior	1	1	0.748	0.002
Scenario 1	1	0.022418	0.785	0.039
Scenario 2	28	1	0.769	0.023
Scenario 3	14.5	0.511209	0.772	0.026

Our heuristic, which chooses $\hat{\alpha}_1$ and $\hat{\alpha}_2$ such that $F(p_L) = \alpha$ works reasonably well in the example with one component having perfect test results, but will likely need to be modified if several components have perfect test results. A large-scale Monte Carlo simulation which involves varying α, the number of system components, the configuration of the system components, and the expected fraction of cases where perfect test results are encountered is the only way to evaluate the techniques presented here, and to compare them, for example, with the asymptotic techniques presented in Mann et al. (1974). Such a simulation is appropriate on a system-by-system basis.

11.5 Simulation

Monte Carlo simulation is used to test several heuristic methods along with the techniques developed in this paper. We begin with a pilot simulation that is used to evaluate a large number of methods in order to thin the number of methods considered.

11.5.1 Simulation Study A

The system considered in this pilot study is a three-component series system with identical components. In keeping with the earlier example, there are n_1=23, n_2=28, and n_3=84 components of each type placed on test. There are B=1000 bootstrap replications used and 1000 simulation replications conducted. The stated coverage of the lower confidence interval bound for the system reliability is 0.95. If the intervals cover approximately 95% of the true system reliability values for a wide range of true component reliabilities, then the confidence interval procedure is performing adequately. Using the two-sided CP confidence interval procedure for a single component described earlier, the acceptable range for the fraction of simulated confidence intervals (at α=0.01) covering the true system reliability is from 0.931 to 0.968 inclusive. The simulations are run in S-Plus using the `set.seed` command prior to each run to exploit the common random numbers variance reduction technique.

In addition, the number of components that achieve perfect test results is computed for general true component reliabilities p_1, p_2, and p_3. For general n_1, n_2, and n_3, let the random variable W be the number of components with perfect test results. The probability density function of W is

$$f(w) = \begin{cases} (1-p_1^{n_1})(1-p_2^{n_2})(1-p_3^{n_3}) & w=0 \\ (1-p_1^{n_1})(1-p_2^{n_2})p_3^{n_3} + (1-p_1^{n_1})p_2^{n_2}(1-p_3^{n_3}) + p_1^{n_1}(1-p_2^{n_2})(1-p_3^{n_3}) & w=1 \\ (1-p_1^{n_1})p_2^{n_2}p_3^{n_3} + p_1^{n_1}(1-p_2^{n_2})p_3^{n_3} + p_1^{n_1}p_2^{n_2}(1-p_3^{n_3}) & w=2 \\ p_1^{n_1}p_2^{n_2}p_3^{n_3} & w=3 \end{cases}$$

These values are computed and given in Table 11.7 for various true, identical component reliabilities ranging from 0.60 to 0.99.

Nine algorithms for handling the case of one or more components with perfect test results are compared in the pilot simulation. We have included algorithms of an ad hoc nature (e.g. Algorithms 2 and 3) and those with some theoretical basis (e.g. Algorithm 9) in order to show that the beta prior approach dominates the other approaches as component reliability increases.

- **Algorithm 1:** Pure bootstrapping. A component with a perfect test always generates perfect simulated results.
- **Algorithm 2:** Always assume a failure. When component i has perfect test results (i.e. $y_i = n_i$), introduce an artificial failure by assuming that $y_i = n_i - 1$, for $i = 1, 2, \ldots, n$.
- **Algorithm 3:** Increase the sample size. For perfect test results, artificially increase sample size to approximate the lower confidence bounds with a single failure using the confidence intervals for a single component given earlier in the chapter, then bootstrap. In our case, $n_1 = 37$ ($y_1 = 36$), $n_2 = 45$ ($y_2 = 44$), and $n_3 = 134$ ($y_3 = 133$).
- **Algorithm 4:** Bayes bootstrapping with $\alpha_1 = 1$ and $\alpha_2 = 1$ (i.e. uniform prior).
- **Algorithm 5:** Bayes bootstrapping with $\alpha_1 = 1$ for all components, and $\alpha_2 = \log(1-\alpha)/\log(1-\alpha^{1/n_i})$, for $i = 1, 2, \ldots, k$, as described earlier.
- **Algorithm 6:** Bayes bootstrapping with $\alpha_1 = n_i$, for $i = 1, 2, \ldots, k$, and $\alpha_2 = 1$ for all components, as described earlier.
- **Algorithm 7:** Bayes bootstrapping with α_1 and α_2 that are averages of the values given in Algorithms 5 and 6.
- **Algorithm 8:** Bayes bootstrapping with $\alpha_1 = 100$ and $\alpha_2 = 1$.
- **Algorithm 9:** A procedure from Mann et al. (1974) which, using asymptotic normal theory, calculates a lower bound as:

$$\prod_{i=1}^{k} \frac{y_i}{n_i} - z_\alpha \sqrt{\prod_{i=1}^{k} \left(\frac{y_i}{n_i}\right)^2 \sum_{j=1}^{k} \left(\frac{1}{y_j} - \frac{1}{n_j}\right)}.$$

The performance of the confidence intervals given in Table 11.7 is as expected. Algorithm 1, for example, which takes the overly optimistic, pure bootstrapping approach, produces lower confidence limits that are shifted up, resulting in fewer-than-expected lower confidence limits that fall below the true system reliability. The opposite case is true for the pessimistic uniform prior distribution in Algorithm 4. In fact, once the Bayesian portion of the algorithm began to dominate (i.e. when the component reliabilities are large), all nine algorithms fail to deliver confidence intervals with the appropriate coverage. We

TABLE 11.7 Estimated Lower Confidence Interval Coverage ($\alpha = 0.05$) for the System Reliability of a Three-component Series System with 1000 Replications using Bootstrapping with a Beta Prior Distribution for Perfect Test Results

True Reliability:	0.60	0.70	0.80	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
Pr(W=0)	0.999	0.999	0.992	0.864	0.822	0.767	0.704	0.621	0.521	0.401	0.267	0.131	0.029
Pr(W=1)	10^{-5}	10^{-4}	0.008	0.132	0.170	0.216	0.271	0.334	0.401	0.460	0.483	0.424	0.221
Pr(W=2)	10^{-12}	10^{-8}	10^{-5}	0.005	0.008	0.014	0.025	0.044	0.078	0.136	0.234	0.380	0.492
Pr(W=3)	10^{-30}	10^{-21}	10^{-13}	10^{-6}	10^{-6}	10^{-5}	10^{-4}	10^{-4}	0.001	0.004	0.016	0.065	0.257
Algorithm 1	**0.958**	**0.964**	**0.946**	0.912	0.910	0.910	0.899	0.899	0.878	0.873	0.809	0.765	0.738
Algorithm 2	**0.958**	**0.964**	**0.945**	**0.950**	**0.960**	0.980	0.988	1.000	1.000	1.000	1.000	1.000	1.000
Algorithm 3	**0.958**	**0.964**	**0.953**	**0.939**	**0.954**	**0.937**	**0.955**	0.972	0.981	1.000	1.000	1.000	1.000
Algorithm 4	**0.958**	**0.964**	**0.948**	**0.960**	**0.957**	**0.968**	0.988	1.000	1.000	1.000	1.000	1.000	1.000
Algorithm 5	**0.958**	**0.964**	**0.939**	0.927	0.918	0.919	0.902	0.910	0.895	0.877	0.832	0.820	0.745
Algorithm 6	**0.958**	**0.964**	**0.947**	0.923	**0.942**	**0.951**	**0.947**	**0.947**	0.970	0.976	1.000	1.000	1.000
Algorithm 7	**0.958**	**0.964**	**0.933**	0.923	**0.947**	0.917	**0.944**	**0.942**	**0.936**	**0.936**	0.982	1.000	1.000
Algorithm 8	**0.958**	**0.964**	**0.951**	0.925	0.925	0.916	0.911	0.926	0.900	0.906	0.876	0.827	1.000
Algorithm 9	0.975	0.947	**0.949**	0.920	0.918	0.919	0.888	0.897	0.886	0.867	0.822	0.737	0.730

The random variable W denotes the number of components with perfect test results. The tabled values give the fraction of intervals that fall below the true system reliability. Fractions set in boldface type are in the range 0.931–0.968 inclusive and are not statistically different from the stated coverage of 0.95.

experimented with the confidence interval that performed the best (Algorithm 7, which averages the parameter estimates of Algorithms 5 and 6) by replacing "average" with "linear combination", but did not produce results that were significantly better than those presented in Table 11.7.

The abysmal performance of all of these algorithms for high-reliability components is consistent with the work of Martz and Duran (1985), who considered lower confidence bounds for the system reliability of 20 system configurations and component reliabilities using three algorithms and two values of α (0.05 and 0.10). Their intervals also diverged from the stated coverages.

11.5.2 Simulation Study B

This poor performance led us to re-code our algorithms in C and to do an exhaustive search in the (α_1, α_2) plane for values of the beta prior parameters α_1 and α_2 that yield reasonable coverages for lower confidence bounds on the system reliability. We returned to the case of a single component. Figure 11.4 shows the results of this exhaustive search for $n=23$ components on test. Every (α_1, α_2) pair that resulted in a confidence interval whose coverage did not statistically differ from 0.95 was plotted for $p=0.91$, $0.92, \ldots, 0.99$. For each particular population reliability p shown in Figure 11.4, the areas where appropriate coverages are achieved are quite narrow. Unfortunately, the graph in Figure 11.4 shows that there is no single (α_1, α_2) pair that will work for all values of p.

The following procedure has been developed as a compromise that allows reasonable lower confidence limit coverage in the case of a system with one or more components having perfect test results:

- For each of the components in the system, consult with someone familiar with the component to get a point estimate of the component reliability p_i^*, $i=1, 2, \ldots, k$.
- Determine the number of components to be tested $n_1, n_2, \ldots, n_k$.
- For each (p_i^*, n_i) pair, perform an exhaustive search of the (α_1, α_2) plane to find a $(\tilde{\alpha}_1, \tilde{\alpha}_2)$ pair which yields appropriate coverage.
- Perform Bayesian bootstrapping as described earlier in this chapter using the test results (n_i, y_i) and the appropriate $(\tilde{\alpha}_1, \tilde{\alpha}_2)$ values from the previous step.

The final example illustrates this technique for a single-component system.

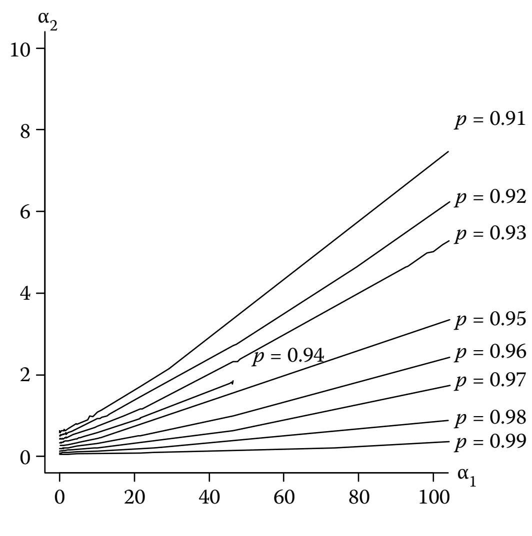


FIGURE 11.4 Prior distribution parameter pairs that give accurate coverage for the example as a function of the reliability p when $n=23$.

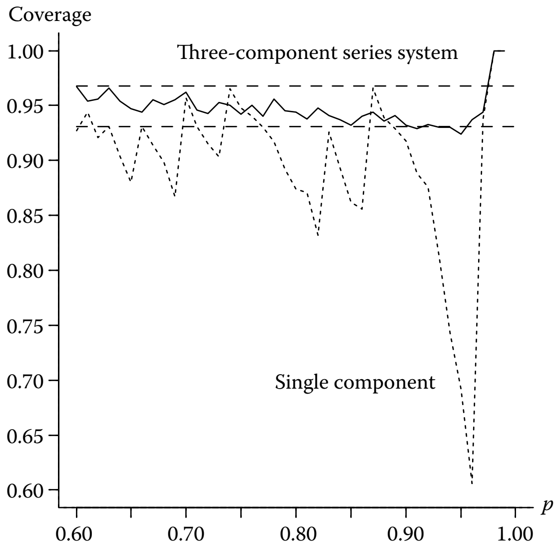


FIGURE 11.5 Lower 95% confidence bound coverage for a single component system (dotted) and a three-component system (solid) and region not statistically different from the specification (dashed).

Example: Single-component System and Three-component Series System

Figure 11.5 plots the actual coverage for 1000 simulation replications for a single-component system with $n=28$ and a three-component system with $n_1=23$, $n_2=28$, and $n_3=84$.

All true component reliabilities are equal, and plotted on the horizontal axis. The stated coverage on all intervals is 0.95. The usual bounds around 0.95 (at 0.931 and 0.968) which denote confidence intervals whose actual coverage does not differ significantly from the specification are given as horizontal dashed lines. All Bayesian procedures use $(\alpha_1, \alpha_2)=(252.28, 4.67)$, which were values that fell outside of the axes in Figure 11.4 associated with the $p^*=0.97$ estimate for the reliability of the second component. The jagged appearance for the coverage for the interval for a single component (dotted) is consistent with the same pattern shown by Blyth (1986). The three-component system (solid), on the other hand, has different numbers of components on test that seem to "average out" these fluctuations, resulting in appropriate coverage through $p=0.97$.

11.6 Conclusion

Determining lower confidence bounds from binary data remains an important yet elusive question. The Bayesian bootstrapping procedures developed here yield adequate coverages given that an expert is able make a good initial estimate of the reliabilities of individual components. The estimates discussed here improve with increasing system complexity.

Acknowledgments

The author gratefully acknowledges the helpful interactions with John Backes, Don Campbell, Mike Crawford, Mark Eaton, Jerry Ellis, Greg Gruver, Jim Scott, Bob Shumaker, Bill Treadwell, and Mark Vinson concerning the chapter and helpful comments from two anonymous referees and the editor.

Appendix 11.1

S-Plus code for calculating a CP (1–α)100% lower confidence bound for a single component for *n* components on test and *y* passes.

```
confintlower<- function(n, y, alpha) {
  if (y==0) {
    pl <- 0
  }
  if (y==n) {
    pl <- alpha ^ (1 / n)
  }
  if (y > 0 && y < n) {
    fcrit1 <- qf(alpha, 2 * y, 2 * (n - y + 1))
    pl <- 1 / (1 + (n - y + 1) / (y * fcrit1))
  }
  pl
}
```

Appendix 11.2

S-Plus code for calculating a bootstrap (1–α)100% lower confidence interval bound for a *k*-component series system of independent components using *B* bootstrap replications.

```
seriessystemboot <- function(n, y, alpha) {
  k <- length(n)
  b <- 10000
  z <- rep(1, b)

  point <- prod(y) / prod(n)

  for (j in 1:b) {
    for (i in 1:k) {
      z[j] <- z[j] * rbinom(1, n[i], y[i] / n[i]) / n[i]
    }
  }
  z <- sort(z)
  pl <- z[floor(alpha * b)]
  c(point, pl)
}
```

Appendix 11.3

APPL code for calculating a bootstrap (1–α)100% lower confidence interval bound for a *k*-component series system of independent components using the equivalent of $B=+\infty$ bootstrap replications.

```
n1 := 23;
y1 := 21;
X1 := BinomialRV(n1, y1 / n1);
X1 := Transform(X1, [[x -> x / n1], [-infinity, infinity]]);
n2 := 28;
y2 := 27;
```

```
X2 := BinomialRV(n2, y2 / n2);
X2 := Transform(X2, [[x -> x / n2], [-infinity, infinity]]);
n3 := 84;
y3 := 82;
X3 := BinomialRV(n3, y3 / n3);
X3 := Transform(X3, [[x -> x / n3], [-infinity, infinity]]);
Temp := Product(X1, X2);
T := Product(Temp, X3);
```

Appendix 11.4

S-Plus code for calculating a bootstrap (1–α)100% lower confidence interval bound for a *k*-component series system of independent component with some perfect component test results using *B* bootstrap replications.

```
seriessystembayesboot <- function(n, y, alpha) {
  k <- length(n)
  alpha1 <- 1
  alpha2 <- 1
  b <- 10000
  z <- rep(1, b)

  point <- prod(y) / prod(n)

  for (j in 1:b) {
    for (i in 1:k) {
      if (y[i] == n[i]) z[j] <- z[j] * rbeta(1, alpha1 + y[i], alpha2)
      else z[j] <- z[j] * rbinom(1, n[i], y[i] / n[i]) / n[i]
    }
  }
  z <- sort(z)
  pl <- z[floor(alpha * b)]
  c(point, pl)
}
```

References

Agresti, A. and Coull, B. 1998. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* 52, 119–126.

Blyth, C.R. 1986. Approximate binomial confidence limits. *Journal of the American Statistical Association* 81, 843–855; Corrigenda (1989), 84, 636.

Chick, S. (2001). Personal communication.

Efron, B. and Tibshirani, R.J. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York, NY.

Evans, M., Hastings, N., and Peacock, B. 2000. *Statistical Distributions*, 3rd edn. John Wiley and Sons, New York, NY.

Glen, A.G., Evans, D.E., and Leemis, L.M. 2001. APPL: A probability programming language. *The American Statistician* 55, 156–166.

Law, A.M. and Kelton, W.D. 2000. *Simulation Modeling and Analysis*, 3rd edn. McGraw-Hill, New York, NY.

Leemis, L.M. and Trivedi, K.S. 1996. A comparison of approximate interval estimators for the Bernoulli parameter. *The American Statistician* 50, 63–68.

Mann, N.R., Schafer, R.E., and Singpurwalla, N.D. 1974. *Methods for Statistical Analysis of Reliability and Life Data*. John Wiley and Sons, New York, NY.

Martin, M.A. 1990. On bootstrap iteration for coverage correction in confidence intervals. *Journal of the American Statistical Association* 85, 1105–1118.

Martz, H.F. and Duran, B.S. 1985. A comparison of three methods for calculating lower confidence limits on system reliability using binomial component data. *IEEE Transactions on Reliability* R-34, 113–120.

Martz, H.F. and Waller, R.A. 1982. *Bayesian Reliability Analysis*. John Wiley and Sons, New York, NY.

Newcombe, R.G. 2001. Logit confidence intervals and the inverse Sinh transformation. *The American Statistician* 55, 200–202.

Padgett, W.J. and Tomlinson, M.A. 2003. Lower confidence bounds for percentiles of Weibull and Birnbaum–Saunders distributions. *Journal of Statistical Computation and Simulation* 73, 429–443.

Vollset, S.E. 1993. Confidence intervals for a binomial proportion. *Statistics in Medicine* 12, 809–824.

12

Assessing the Reliability of a Contingency Logistics Network

Marlin U. Thomas
Air Force Institute of Technology

12.1 Introduction

Contingencies are unexpected crises events that create a major threat to the safety and security of a population. They range from military conflicts that require engagement with hostile forces, police actions for civil disorder, to humanitarian relief of victims from disasters such as earthquakes, hurricanes, floods, and related catastrophes. Hurricane Mitchell, the overthrow of the government of Somalia, and the 1990 invasion of Kuwait are examples of crises events that resulted in major contingency operations. Contingency events trigger immediate need for logistics support functions to organize and mobilize people, equipment, materials and supplies for carrying out the emergency operations. These operations are generally initiated from existing plans by the Federal Emergency Management Agency (FEMA) for disasters that occur within the US continent, or the Department of Defense for national contingencies outside the continent.

The logistics support functions for contingencies can be structured as a network of supply chains for receiving, transporting, and distributing materials and equipment for ensuring that "the right things are provided at the right place at the right time". A fairly extensive literature on supply chain management has evolved over the past 15 years (see Tayur et al. 1999). Most of this work has focused on business planning decisions on the distribution, allocation and maintenance of inventories for an assumed infinite time horizon. Supply chains for contingency operations are complicated by the dynamics, risk and uncertainties of the operations, which make it particularly difficult to establish measures of effectiveness for the logistics functions. For business decisions the criteria is generally to maximize profits or seek cost reductions so the measures of supply chain effectiveness are usually based on minimum cost. Contingency operations emerge from crises conditions and time often is a more critical variable than cost in making decisions, particularly early in the deployment of logistics support for the operations. Therefore the likelihood of having the right supplies as needed is a more significant measure of effectiveness for contingency logistics systems (CLS). With the possible exception of Thomas (2002), there appears to be no reported results for quantifying logistics effectiveness for contingency operations.

In this chapter, we describe a method for quantifying the effectiveness of a CLS based on logistics network reliability. Following our notation, in the next section we develop a model for determining the CLS reliability by treating each site as a supply chain link. The logistics network reliability will depend among other things on the conditions for failure of these chain links. Reliability measures are developed for the case of a three-site CLS for various conditions of risk. Results are provided for known supply and demand distributions at each supply support site. Sometimes it is impractical to treat the demand distribution as known but some information can be assumed through moments or event probabilities. For these cases, maximum entropy (ME) distributions are derived and used for the demand distribution. Some concluding remarks are given in the final section.

12.2 Notation

X_j	Random variable representing the supply at site $j=1, \ldots, n$
Y_j	Random variable representing the demand at site j
F_j	Probability distribution of the supply at site j
G_j	Probability distribution of the demand at site j
g_{ME}	Maximum entropy probability density function
μ_i	i^{th} moment about the origin
Ω	Sample space containing all outcomes for the supply chain
ω_{ij}	Outcome i for supply site j where $i=1, \ldots, n$
p_j	Failure probability of supply site j, $0<p_j<1$
q_j	Success probability of supply site j, $q_j=1-p_j$
R_S	System reliability

12.3 CLS

We define a CLS as a set of processes and methods for providing the procurement, distribution, storage and transportation of people, supplies, materials, and equipment for supporting contingency operations. While contingencies evolve from random events there are plans that provide guidance on the logistics required, at least for the mobilization and build-up of the forces needed to accomplish the mission. The life cycle and relative resource profile for a contingency is illustrated by the ramp function in Figure 12.1.

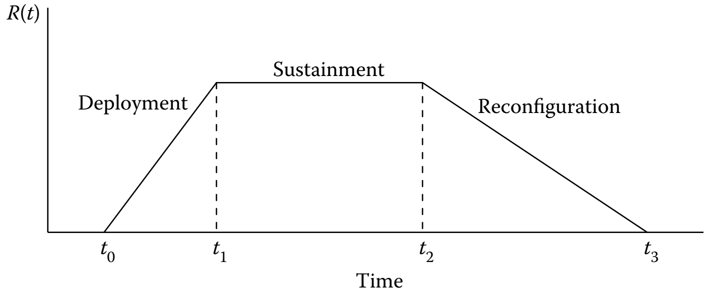


FIGURE 12.1 Contingency life-cycle resource loading profile.

There are three distinct phases of the life cycle: deployment, sustainment, and reconfiguration. The deployment phase starts at a time t_0, marking the beginning of the contingency and extends throughout the mobilization and build-up to a time t_1. As the force or organization builds up, the logistics requirements also ramp up. Once build-up is completed, the operations are sustained throughout the period (t_1,t_2). During this phase the nature of the operation and logistics requirements can change with the tempo and operational conditions but the basic processes and structure remain in place. The length of the sustainment period could be weeks, months, or years, but at some point, t_2, the operation is terminated either by completion of the mission or by directive of higher authority. During this reconfiguration phase service centers for supporting the mobilized personnel and equipment are phased down or collapsed as necessary to accommodate declining demands, as redeployment occurs. Supplies and assets, including personnel and equipment are redistributed and returned.

12.4 Reliability Model Formulation

Operationally, a CLS can be viewed as a network of supply chains for routing the right logistics items to the right place in support of the contingency. Consider a system where supplies are received at a single distribution center where they are allocated among *n* support sites, as shown in Figure 12.2. This could be a distribution center for a critical supply element such as the ammunition for a military contingency in a hostile environment. All ammunition is transported to the distribution center, which is at a higher echelon level and provides the command and control support to the *n* logistics sites. We further assume that the demand at each support site is independent of the on-hand inventory.

The method for determining the CLS reliability is to: (1) establish the conditions for failure at each link of the chain; (2) determine the link failure probabilities; and (3) compute the system reliability for the prescribed system configuration.

12.4.1 Failure Conditions and Event Probabilities

Let X_j and Y_j be independent random variables with probability distributions $F_j(x)$ and $G_j(y)$ representing the supply and demand at each site $i=1, \ldots, n$. Applying concepts from interference theory (see Kapur and Lamberson 1977), we define a failure by the interference event $\{X_j < Y_j\}$ where the on-hand inventory of mission critical materials is exceeded by the demand. Denoting the outcomes at each site by

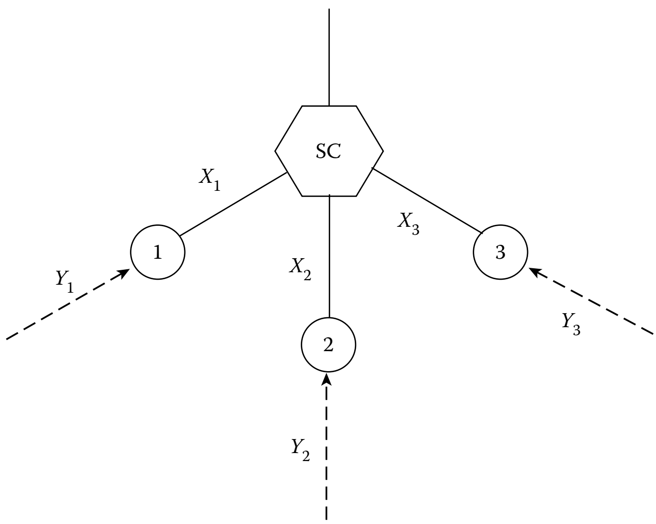


FIGURE 12.2 A supply distribution chain for a system with three sites.

$$\omega_i = \begin{cases} 0, & X_i < Y_i \\ 1, & X_i \geq Y_i \end{cases}, \tag{12.1}$$

the sample space for the operational readiness of the network,

$$\Omega = \{(\omega_1, \ldots, \omega_n)_j; j = 1, \ldots, 2^n\} \tag{12.2}$$

contains 2^n n-tuples. For the case of $n=2$, $\Omega = \{(\omega_1, \omega_2)_j; j=1, \ldots, 4\}$ which is the set of sample points: (1,1), (1,0), (0,1), (0,0).

It follows by conditioning on Y_i that the probability of failure at site i is given by

$$p_i = \begin{cases} \sum_{y=0}^{\infty} F_i(y) g_i(y), & \text{for discrete}\,(X_i, Y_i) \\ \int_0^{\infty} F_i(y) dG_i(y), & \text{for continuous}\,(X_i, Y_i) \end{cases} \tag{12.3}$$

12.5 A CLS with Three Support Sites

For purposes of development we consider a CLS with three logistics sites. The results that follow are easily extended to a more general system with n sites. Consider the CLS in Figure 12.2. Since we have $n=3$ logistics sites, the sample space from Equation 12.2 is

$$\Omega = \{(\omega_1, \omega_2, \omega_3)_j; j = 1, \ldots, 8\} \tag{12.4}$$

The procedure for estimating the CLS reliability is summarized as follows:

1. Develop the system configuration by establishing the criticality of each site relative to its dependency in complying with the requirements of the mission.
2. Establish the failure conditions; identifying the failure modes for each site, as linkages in the chain.
3. Estimate the failure probabilities by assuming the distributions and parameter values for the supply and demand distributions at each site using the best information available.
4. Compute the reliability R for the given failure conditions established in 1 and 2 above.

The system failure criteria will vary with the type of system and criticality of each link relative to the particular mission. Under extreme conditions each link relates to a unique function that is an integral part of the operation. This corresponds to a series configuration where the failure of any link constitutes failure of the system. The other extreme case is a parallel configuration where any functional link can support the demand requirements for the operation. The k-out-of-n configuration is another option whereby there is some minimal requirement for k of the logistics sites to be functional in order to have an operational system. The system failure events for these conditions are given in Table 12.1.

12.6 Models for Known Supply and Demand Distributions

For cases where from past contingencies and related scenarios the distributions for the supply and demand can be treated as known the failure probability can be determined from Equation 12.3. To illustrate consider the following two examples for computing link failure probabilities.

TABLE 12.1 The Link Failure Conditions for a CLS with Three Support Sites

			System Failure		
j	$(\omega_1, \omega_2, \omega_3)_j$	$P\{(\omega_1, \omega_2, \omega_3)_j\}$	Series	Parallel	2-of-3
1	(1,1,1)	$q_1q_2q_3$			
2	(1,1,0)	$q_1q_2p_3$	x		
3	(1,0,1)	$q_1p_2q_3$	x		
4	(0,1,1)	$p_1q_2q_3$	x		
5	(1,0,0)	$q_1p_2p_3$	x		x
6	(0,1,0)	$p_1q_2p_3$	x		x
7	(0,0,1)	$p_1p_2q_3$	x		x
8	(0,0,0)	$p_1p_2p_3$	x	x	x

12.6.1 Erlang Supply and Exponential Demand

Consider a given site i with the supply and demand probability density functions given by

$$f_{X_i}(x) = \frac{\lambda_i(\lambda_i x)^{r_i-1}}{(r_i-1)!}e^{-\lambda_i x},\ \lambda_i > 0,\ r_i = 1, 2, \ldots;\ x \geq 0 \tag{12.5}$$

and

$$g_{Y_i}(y) = \mu e^{-\mu y},\ \mu > 0;\ y \geq 0 \tag{12.6}$$

To determine the cumulative distribution,

$$F_{X_i}(y) = \int_0^y \frac{\lambda_i(\lambda_i x)^{r_i-1}}{(r_i-1)!}e^{-\lambda_i x}\mathrm{d}x$$

letting $w=\lambda_i x$ and hence $\mathrm{d}w=\mathrm{d}x/\lambda_i$, it follows that

$$F_{X_i}(y) = \int_0^{\lambda_i y} \frac{w^{r_i-1}}{(r_i-1)!}e^{-w}\mathrm{d}w = \left[1 - \left(1 + \lambda_i y + \frac{(\lambda_i y)^2}{2!} + \ldots + \frac{(\lambda_i y)^{r_i-1}}{(r_i-1)!}\right)e^{-\lambda_i y}\right] \tag{12.7}$$

Substituting Equation 12.7 into Equation 12.3 with $g_{Y_j}(y)$ from Equation 12.6 leads to the failure probability

$$p_i = \left(\frac{\lambda_i}{\lambda_i + \mu_i}\right)^{r_i+1} \tag{12.8}$$

Example 12.1

A contingency support force is deployed to a remote location to provide emergency assistance and medical relief to victims of a major hurricane. Logistics support, which includes medical supplies, is distributed through three regional sites. A particular serum is required for treating the selected victims of the hurricane, which is received at each site in accordance with an Erlang distribution with parameters $\lambda=0.01$ and $r=2$. The demand for the medicine at each site is distributed exponential with a mean of 50 units. The distribution chain is assumed to be functional so long as the serum is available at one of the sites.

So here we have $\lambda_1=\lambda_2=\lambda_3=0.01$, $r_i=2$, and $\mu_1=\mu_2=\mu_3=0.02$, and in Equation 12.8 the site failure probabilities are

$$p_j=\left(\frac{0.01}{0.01+0.02}\right)^{2+1}=\frac{1}{27}$$

For an assumed parallel system from Equation 12.4, the reliability is

$$R_S=1-p_1p_2p_3=1-\left(\frac{1}{27}\right)^3=0.99995$$

12.6.2 Normal Supply and Demand

For the case where the supply $X\sim N(\mu_X,\sigma_X^2)$ and the demand $Y\sim N(\mu_Y,\sigma_Y^2)$,

$$f_{X_j}(x)=\frac{1}{\sigma_{X_j}\sqrt{2\pi}}e^{-\frac{(x-\mu_{X_j})^2}{2\sigma_{X_j}^2}},\ \mu_{X_j}\geq 0,\ \sigma_{X_j}>0,\ -\infty<x<\infty \tag{12.9}$$

and

$$g_{Y_j}(y)=\frac{1}{\sigma_{Y_j}\sqrt{2\pi}}e^{-\frac{(y-\mu_{Y_j})^2}{2\sigma_{Y_j}^2}},\ \mu_{Y_j}\geq 0,\ \sigma_{Y_j}>0,\ -\infty<y<\infty$$

therefore,

$$F_{X_j}(y)=\Phi\left(\frac{y-\mu_{Y_j}}{\sigma_{Y_j}}\right). \tag{12.10}$$

The failure probabilities for each site are given by

$$p_j=\Phi\left(-\frac{\mu_{X_j}-\mu_{Y_j}}{\sqrt{\sigma_{X_j}^2+\sigma_{Y_j}^2}}\right). \tag{12.11}$$

To prove this rather than deriving it directly from Equation 12.2, it is easier to determine the distribution of $W_j=X_j-Y_j$ and then find $H_W(0)=P(W\leq 0)$. The Laplace transforms for the $f_{X_j}(x)$ and $g_{Y_j}(y)$ are

$$\bar{f}_{X_j}(s) = e^{-\mu_{Xj}s + \sigma_{X_j}^2 s^2/2}$$

and

$$\bar{g}_{Y_j}(s) = e^{-\mu_{Y_j}s + \sigma_{Y_j}^2 s^2/2},$$

respectively, and since X_j and Y_j are independent the transform for the probability density function for W_j is

$$\bar{h}_W(s) = E\left(e^{-s(X_j - Y_j)}\right) = \left(e^{-\mu_{Xj}s + \sigma_{X_j}^2 s^2/2}\right)\left(e^{-\mu_{Y_j}(-s) + \sigma_{Y_j}^2 (-s)^2/2}\right),$$

or

$$\bar{h}_W(s) = e^{-\left(\mu_{Xj} - \mu_{Y_j}\right)s + \left(\sigma_{X_j}^2 + \sigma_{Y_j}^2\right)s^2/2},$$

which implies that W is distributed $N(\mu_{X_j} - \mu_{Y_j}, \sigma_{X_j}^2 + \sigma_{Y_j}^2)$. The failure probability for site j is then

$$p_j = P(W \le 0) = \Phi\left(\frac{-\left(\mu_{X_j} - \mu_{Y_j}\right)}{\sqrt{\sigma_{X_j}^2 + \sigma_{Y_j}^2}}\right). \tag{12.12}$$

Results for computing the link probabilities for several supply and demand distributions are given in Table 12.2. For more details on the reliability interference concept the reader is directed to Kapur and Lamberson (1977).

12.7 ME Demand

Contingency operations by their nature can be difficult due to uncertainties involved, particularly during the sustainment period where the duration, demands, and required supplies can all be uncertain. We consider now the situation where the distribution of supplies are known but only limited information is known about the demands. The ME principle provides a rational basis for determining a distribution based on minimal but factual information about its moments and selected event probabilities, Jaynes (1957). For a continuous nonnegative random variable Y, the ME probability density function, $g_{\mathrm{ME}}(y)$ is determined by solving the optimization problem,

$$\begin{aligned}
&\max_g H(g) = -\int g(y)\ln g(y)\,\mathrm{d}y \\
&\text{st:} \int_0^\infty g(y)\,\mathrm{d}y = 1,\ g(y) \ge 0 \\
&\int \xi_k(y) g(y)\,\mathrm{d}y = \mu_k;\ k = 1, 2, \ldots
\end{aligned} \tag{12.13}$$

TABLE 12.2 Link Failure Probabilities for Selected Distributions

Demand, $f(x)$	Supply, $g(y)$	Failure probability, $0<p<1$
$\alpha(1-\alpha)^x,\ x=0,1,\ldots$	$\beta(1-\beta)^y,\ y=0,1,\ldots$	$\dfrac{\alpha}{\alpha+\beta-\alpha\beta}$
$\lambda e^{-\lambda x},\ x\geq 0$	$\mu e^{-\mu y},\ y\geq 0$	$\dfrac{\lambda}{\lambda+\mu}$
$\dfrac{\lambda(\lambda x)^{r-1}}{(r-1)!}e^{-\lambda x},\ r=1,2,\ldots;\ x\geq 0$	$\mu e^{-\mu y},\ y\geq 0$	$\left(\dfrac{\lambda}{\lambda+\mu}\right)^r$
$\dfrac{\lambda(\lambda x)^{r-1}}{\Gamma(r)}e^{-\lambda x},\ x\geq 0$	$\dfrac{\mu(\mu y)^{s-1}}{\Gamma(s)}e^{-\mu y},\ y\geq 0$	$\dfrac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}B_{\frac{\mu}{\lambda+\mu}}(r,s)$
$\dfrac{e^{-\left(\frac{x-\mu_X}{\sigma}\right)^2}}{\sigma_X\sqrt{2\pi}},\ -\infty<x<\infty$	$\dfrac{e^{-\left(\frac{y-\mu_Y}{\sigma}\right)^2}}{\sigma_Y\sqrt{2\pi}},\ -\infty<y<\infty$	$\Phi\left(-\dfrac{\mu_X-\mu_Y}{\sqrt{\sigma_X^2+\sigma_Y^2}}\right)$
$\lambda e^{-\lambda x},\ x\geq 0$	$\dfrac{1}{\sigma_Y\sqrt{2\pi}}e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}},\ -\infty<y<\infty$	$\Phi\left(-\dfrac{\mu_X}{\sigma_X}\right)+\left[1-\Phi\left(-\dfrac{\mu_X-\lambda^2\sigma_X^2}{\sigma_X}\right)\right]e^{-\frac{\lambda}{2}(2\mu_X-\lambda\sigma_X^2)}$

where $H(g)$ is the entropy associated with the random variable Y and $\xi_k(y)$ is an integrable function of y with μ_k constant for all values of k. This problem will always have as a minimum, with the constraints that $g(y)$ is nonnegative and integrates to 1. The remaining constraints are based on the known input from the application environment. The solution to Equation 12.13 can be found through a conventional technique such as the calculus of variations to obtain

$$g_{ME}(y)=e^{-\pi_0-\sum\pi_k\xi_k(y)},\ k=1,2,\ldots;\ y\geq 0 \tag{12.14}$$

where the coefficients π_i are determined from the constraints in Equation 12.13. To illustrate this method let us consider the following example.

Example 12.2

Consider the contingency support force in Example 12.1 that distributes the medical serum to hurricane victims through three regional support sites. Suppose the serum is supplied to each site according to a known distribution. The actual demand distributions at the sites are not known but some information on the demand can be contrived from intelligence and past experiences that allows for the following assumed conditions.

12.7.1 Normal Distributed Supply and Known Mean Demand

Suppose the serum is supplied at each site according to a normal distribution with mean μ_X and variance σ_X^2. The only information available on the demand at a given site is through an assumed mean value μ_1. So letting $\xi_1(y)=\mu_1$ in Equation 12.13 the ME density is given by:

$$g_{\mathrm{ME}}(y)=e^{-\pi_0+\pi_1 y},\quad y\geq 0$$
$$\text{st}:\ \int_0^{\infty} g_{\mathrm{ME}}(y)\mathrm{d}y=1, \tag{12.15}$$
$$\int_0^{\infty} y g_{\mathrm{ME}}(y)\mathrm{d}y=\mu_1$$

with $g_{\mathrm{ME}}(y)\geq 0$. Substituting $g_{\mathrm{ME}}(y)$ into the constraints and solving for the coefficients π_0 and π_1, it follows that the resulting ME distribution is the exponential

$$g_{\mathrm{ME}}(y)=\frac{1}{\mu_1}e^{-y/\mu_1},\quad y\geq 0. \tag{12.16}$$

This result has intuitive appeal since the exponential distribution, through its lack of memory property represents the most uncertain conditions when only the mean is known for a continuous random variable.

Substituting g_{ME}(y) into Equation 12.3 it can be shown that the failure probability for a site is

$$p=\Phi\left(-\frac{\mu_X}{\sigma_X}\right)+\left[1-\Phi\left(-\frac{\mu_X-(\sigma_X/\mu_1)^2}{\sigma_X}\right)\right]e^{-\frac{1}{2\mu_1}\left(2\mu_X-\sigma_X^2/\mu_1\right)} \tag{12.17}$$

which is given in Table 12.2 (with $\lambda=1/\mu_1$).

Suppose the means for the three sites are $\mu_{1,1}=100$, $\mu_{1,2}=75$, and $\mu_{1,3}=150$ and the normally distributed supply at each site has a mean of 200 units and variance of 20,000 units^2. Then substituting into Equation 12.17 the failure probability for site 1 is

$$p_1=\Phi\left(-\frac{200}{141.4}\right)+\left[1-\Phi\left(\frac{200-[141.4/100]^2}{141.4}\right)\right]e^{-\frac{1}{2(100)}\left[2(200)-\frac{20{,}000}{100}\right]}=0.63623.$$

Similarly, for sites 2 and 3, $p_2=0.45593$ and $p_3=0.45709$. For a two-out-of-three system configuration, the reliability from Table 12.1 is

$$R_s=q_1q_2q_3+q_1q_2p_3+q_1p_2q_3+p_1q_2q_3$$

or,

$$R_s=(0.36377)(0.54407)(0.54291)+(0.36377)(0.54407)(0.45709)+(0.36377)(0.45593)(0.54291)$$
$$+(0.63623)(0.54407)(0.54291)=0.45789$$

12.7.2 Normally Distributed Supply and Known Mean and Variance of Demand

Suppose now that at a given site in addition to the mean μ_1 being known, the second moment μ_2 is also known for the demand. Thus, we have $\xi_1(y)=y$, $\xi_2(y)=y^2$ in Equation 12.14 and the ME probability density function for the demand Y is given by

$$g_{\mathrm{ME}}(y)=e^{-\pi_0-\pi_1 y-\pi_2 y^2} \tag{12.18}$$

with constraints

$$\int_0^{\infty} g_{\mathrm{ME}}(y)\mathrm{d}y = 1$$

$$\int_0^{\infty} y g_{\mathrm{ME}}(y)\mathrm{d}y = \mu_1$$

$$\int_0^{\infty} y^2 g_{\mathrm{ME}}(y)\mathrm{d}y = \mu_2$$

with $g_{\mathrm{ME}}(y) \geq 0$. Substituting $g_{\mathrm{ME}}(y)$ into the constraints, it can be shown that the resulting ME demand probability density function is the normal

$$g_{\mathrm{ME}}(y) = \frac{e^{-\left(\frac{y-\mu_1}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}, \quad -\infty < y < \infty. \tag{12.19}$$

For this case the failure probability for a site is given in Table 12.2, replacing parameters μ_{X_j} and $\sigma^2_{X_j}$ by the assumed respective values $\mu_{1,j}$ and σ^2_j, noting that $\sigma^2_j = \mu_{2,j} - \mu^2_{1,j}$, for $j=1,\ldots,n$. Suppose the known means and variances are $\mu_{1,1}=100$, $\sigma^2_1 = 10{,}000$, $\mu_{1,2}=75$, $\sigma^2_2 = 10{,}000$, and $\mu_{1,3} = 150$, $\sigma^2_3 = 25{,}000$. From Equation 12.12 the failure probability for link 1 is

$$p_1 = \Phi\left(-\frac{(200-100)}{\sqrt{20{,}000+10{,}000}}\right) = \Phi(-0.577) = 0.282$$

and similarly, $p_2=0.485$ and $p_3=0.407$. For a two-out-of-three-system configuration, the reliability from Table 12.1,

$$R_S = q_1q_2q_3 + q_1q_2p_3 + q_1p_2q_3 + p_1q_2q_3$$

or

$$R_S = (0.717)(0.515)(0.593) + (0.717)(0.515)(0.407) + (0.717)(0.485)(0.593) + (0.282)(0.515)(0.407) = 0.6346$$

12.7.3 Exponential Distributed Supply and ME Distributed Demand

While a wide range of planning assumptions can be incorporated in Equation 12.13 to establish ME distributions for contingency logistics demand, we note that it would be extremely rare for the knowledge about Y to involve more than two moments. It is reasonable however to construct events of the form

$$P(Y > L) \leq \gamma$$

from existing physical conditions or through intelligence data. Incorporating this condition along with a known mean for the demand Y, the ME problem of Equation 12.13 becomes

$$\max_{g} H(g) = -\int_{0}^{\infty} g(y)\ln g(y)\mathrm{d}y$$

$$\mathrm{st:}\quad \int_{0}^{\infty} g(y)\mathrm{d}y = 1$$

$$\int_{0}^{\infty} g(y)\mathrm{d}y \le \gamma_L \tag{12.20}$$

$$\int_{0}^{\infty} y\,g(y)\mathrm{d}y = \mu_1$$

$$g(y) \ge 0,\ y \ge 0$$

It follows that the solution to Equation 12.20 is

$$g_{\mathrm{ME}}(y) = \begin{cases} (\gamma_L/\mu)e^{-y/\mu_1}, L < y < \infty \\ \dfrac{1-\gamma_L e^{-L/\mu_1}}{\mu\left(1-e^{-L/\mu_1}\right)} e^{-y/\mu_1}, 0 \le y \le L \end{cases} \tag{12.21}$$

and substituting into Equation 12.3, the site failure probability is given by

$$p = \alpha_1 \int_{0}^{L} F_X(y)e^{-y/\mu_1}\mathrm{d}y + \alpha_2 \int_{L}^{\infty} F_X(y)e^{-y/\mu_1}\mathrm{d}y \tag{12.22}$$

where

$$\alpha_1 = \frac{1-\gamma_L e^{L/\mu_1}}{\mu_1\left(1-e^{-L/\mu_1}\right)},\ \text{and}\ \alpha_2 = \gamma_L/\mu_1 .$$

For the case of ME distributed demand given by Equation 12.21 and supply X distributed exponential with

$$F_X(x) = 1 - e^{-\lambda x},\ x \ge 0,$$

it follows, adding the additional subscript j that the link failure probability for a site $j = 1,2,3$ is given by

$$p_j = 1 - \left(\frac{1}{\lambda_j + 1/\mu_{1,j}}\right)\left\{\alpha_{1,j} + \left(\alpha_{2,j} - \alpha_{1,j}\right)e^{-L_j(\lambda_j + \mu_{1,j})}\right\} \tag{12.23}$$

For illustration let us return to Example 2 but now we assume that supply of the serum at each site is distributed exponentially with $\lambda = 5\times10^{-5}$. The mean demand at the sites is $\mu_{1,1} = 100$, $\mu_{1,2} = 75$, and $\mu_{1,3} = 150$ units. We further assume that the probabilities of the demand exceeding level $L_1 = 50$ units at site 1, $L_2 = 25$ units at site 2, and $L_3 = 75$ units at site 3 are each bounded by $\gamma_L = 0.75$. Substituting these values into Equation 12.22 we determine the coefficients $\alpha_{1,j}$ and $\alpha_{2,j}$ for each of the sites $j = 1,2,3$ as follows

$$\alpha_{1,1} = 0.013854,\ \alpha_{2,1} = 0.0075$$

$$\alpha_{1,2} = 0.021759,\ \alpha_{2,2} = 0.01,$$

$$\alpha_{1,3} = 0.0098707,\ \alpha_{3,3} = 0.005,$$

and from Equation 12.23, the failure probabilities are found to be $p_1 = 0.00402$, $p_2 = 0.00295$, and $p_3 = 0.00601$. So if each site is required to function without back-up from the other sites, the reliability for the system is

$$R_s = q_1 q_2 q_3 = (0.99598)(0.99705)(0.99399) = 0.98707.$$

12.8 Concluding Remarks

The method described in this chapter provides a means for assessing the effectiveness of a contingency logistics system, or any supply chain network comprised of well-defined links and nodes associated with random supply and demand. Here we considered a single stage process for a contingency with a mission of finite duration, and known distributions for supply and demand. Data is most always quite limited or nonexistent for the demand in a CLS though some knowledge is generally available. The challenge is to rationally incorporate this knowledge in selecting the distributions and parameter values for making decisions in allocating the logistics resources.

The definitions for the failure conditions are of course very important in quantifying the reliability. The same scrutiny must be applied to the logistics functions that are applied in conducting a failure mode and criticality analysis for a product development. The reliability of the contingency logistics system will depend upon the particular configuration of the supply chain relative to the sharing and networking rules. In some cases one supply site might share with another site when failure occurs, whereas in other cases this might be impractical. For example, in a tense and hostile environment where the supply of a critical resource such as ammunition is low, the higher command will normally order a reallocation that includes the return and redistribution of supplies from units to provide a balance. The mission and environment of the particular contingency operations will determine the conditions for supply chain failure and overall reliability.

The directions for further developments are suggested in two areas. Since contingency operations are characterized by uncertainties as well as the dynamics of the operations, there is need for a method of incorporating risk in establishing rational decision criteria for making CLS decisions. One approach is to apply ME concepts described in the previous section. This has a rational appeal in reliability and for practical applications such as treating the uncertainties of failure conditions for a supply chain link where knowledge about a distribution other than its moments is often available. A generalized procedure can be applied using ME concepts for broad conditions on an underlying distribution (Thomas 1979). Other approaches for incorporating risk and uncertainty should also be explored.

A second area important for managing and controlling logistics for contingency operations is the development of allocation rules for distributing critical resources. When a supply item like fuel, ammunition, or water is limited in quantity but is critical to the success of the mission it is rationed among the n logistics sites. The problem is to find an allocation of the available resource for each site. Without particular knowledge of the future, one approach is to simply allocate uniformly across all sites. This presumes that each site requires equal demand and is of equal importance in accomplishing the overall mission. Various criteria can be applied and should be examined relative to the types of information that is available for contingencies. The criteria can be different depending upon the type of contingency and whether it is in the deployment, sustainment, or reconfiguration stage.

Currently, there are no known quantitative methods that are being applied in evaluating contingency plans. The method for assessing CLS reliability described in this paper provides a means for evaluating alternative plans, mission scenarios and associated risks for contingencies. The results are also useful in developing continuous quality improvement goals for contingency logistics systems.

Acknowledgment

The author is grateful to the National Science Foundation for the support of this work under Grant 0071935-DMI.

References

Jaynes, E.T. 1957. Information theory and statistical mechanics. *Physics Review*, 106, 620–630.

Kapur, K.C. and L.R. Lamberson. 1977. *Reliability in Engineering Design.* John Wiley & Sons. New York.

Tayur, S., R. Ganeshan, and M. Magazine (Eds). 1999. *Quantative Models for Supply Management.* Kluwer Academic Publishers. Norwell, MA.

Thomas, M.U. 1979. A generalized maximum entropy principle. *Operations Research*, 27 (6), 1198–1196.

Thomas, M.U. 2002. A method for assessing supply chain reliability for contingency operations. Research Memorandum, No. 02-02, School of Industrial Engineering, Purdue University, West Lafayette, IN.

13

Computing Small-Fleet Aircraft Availabilities Including Redundancy and Spares

Jeffery K. Cochran
Air Force Institute of Technology

Theodore P. Lewis
Air Force Material Command Headquarters

13.1 Background

The United States Air Force currently uses the best analytical model available for the aircraft spares provisioning problem. It is named the Dynamic Multi-Echelon Technique for Recoverable Item Control (Dyna-METRIC) developed by the RAND Corporation [1]. Dyna-METRIC is a flexible inventory modeling tool, but its underlying assumption is that a large number of entities are being modeled. The following is a brief discussion of the Dyna-METRIC model. The central theorem in the Dyna-METRIC model is Palm's theorem [2], also known as the infinite channel queuing assumption [3], which states that if demand for an item is given by a Poisson process with mean m per unit time, and if the repair time for each failed unit is independently and identically distributed according to any distribution with mean repair time T, then the steady-state probability distribution for the number of units in repair has a Poisson distribution with mean mT.

This is a powerful result since it does not depend on any assumption about underlying probability distributions, but when dealing with a small number of aircraft this theorem is violated because the demand distribution and the repair cycle are no longer independent. The problem is well-documented and can occur with as many as eight aircraft [4].

The number and location of s spare parts in a supply system is:

$$s = \text{OH} + \text{DI} - \text{BO} \tag{13.1}$$

where OH, the number of spares on the shelf or on-hand; DI, the number of items due-in from repair or resupply; BO, the number of items on backorder.

This is a balance equation where the order quantity is assumed to be one. Any change in one variable will result in a compensating change in another variable. For example, if an item demand occurs, the number of items due-in will increase by one and, if the current on-hand balance is greater than zero, then the on-hand balance will decrease by one. Otherwise, the number of backorders will increase by one.

Using the initial spares position of each item type i, the Dyna-METRIC model predicts the number of aircraft available and the number of sorties that can be flown each day. In order to make these predictions, the model must calculate the expected number of backorders $\mathrm{EBO}_i(s_i)$ for each item type on the aircraft. This is done using:

$$\mathrm{EBO}_i(s_i) = x = \sum_{s_i+1}^{\infty} (x - s_i)\mathrm{Pr}_i[\mathrm{DI} = x] \tag{13.2}$$

where s_i, the stock level for each item type i; $\mathrm{Pr}_i[\mathrm{DI}]$, probability of DI items due in from repair or resupply for item type i.

An important common logistics measurement is "pipeline". Pipeline represents the number of units of an item in the repair or in the resupply chain. The average pipeline, μ, for the single base, full repair, no depot resupply case is the average demand, m, multiplied by the repair time, T, such that $\mu = mT$. As a result of Palm's Theorem, this average pipeline value becomes the mean of the Poisson distribution used to calculate the expected backorders. If we allow multiple bases, limited base repair, and depot repair and resupply, the average pipeline at base j becomes:

$$\mu_j = m_j(r_jT_j + (1 - r_j)[O_j + \mathrm{EBO}[s_o|m_oT_o]/m_o]) \tag{13.3}$$

where μ_j, average pipeline at base j; m_j = average annual demand at base j; r_j = probability of repair at base j; T_j, average repair time at base j; O_j = average order and ship time from depot to base j; subscript j = base counter; subscript o, depot counter.

For most single-unit aircraft combat assessments, this equation reduces to $\mu_j = (r_jT_j)m_j$ since depot repair and resupply are not available.

Aircraft availability (the ratio of aircraft available to aircraft fielded times 100) is an important measure of merit [5]. It can be calculated from the expected backorders as follows:

$$A = 100\prod_{i=1}^{I}[1 - \mathrm{EBO}_i(s_i)/(NZ_i)]^{Z_i} \tag{13.4}$$

Subject to $\mathrm{EBO}_i(s_i) \le NZ_i$ for every item type i

where Z_i, number of times the same item occurs on a single aircraft; N, number of aircraft fielded.

This formula implies that an aircraft is available only when there are no failures in any of the Z_i items on an aircraft. The constraint prevents the number of backorders from exceeding the number of possible aircraft positions for each item type. The number of predicted aircraft flights/flying hours per day is simply the number of available aircraft each day times the maximum flight rate (flights/aircraft) per day which is capped at the total number of flights required each day.

This is a brief overview of the current model and a description of the current state of the modeling environment. Next, we introduce a method that addresses some of the shortcomings of the current method.

13.2 Research Method

The following method is an effort to better model small systems of equipment where some of the assumptions made in Dyna-METRIC fail. However, the proposed method still maintains the advantages of the analytical model. The proposed method must overcome several challenges. These include the difficulties of a finite-calling population, the proper distribution of component backorders, item redundancy, and the interdependence of component types. Previous researchers have dealt with the first two of these. We begin with these, then extend that redundancy work. Subsequently, we present a new iterative approach to handling interdependence.

Below is a failure distribution function from queuing theory that takes into account the finite calling nature of our situation [6]. Once all the original operating units and spares are consumed, no more arrivals (or failures) can occur. Also, as the number of operating components decreases, the arrival rate or failure rate naturally decreases as well. This is the probability distribution of failures, $P(X)$:

$$P(X=x)=P(0)\frac{f^xR^x}{x!}\text{ when } x\leq s,\text{ and } P(0)\frac{f^xR^s}{x!}\frac{R!}{(R-x+s)!}\text{ when } s+1\leq x\leq R+s \tag{13.5}$$

where x, number of component failures; f, ratio of failure rate to repair rate (λ/μ); $P(0)$, normalizing constant so the distribution sums to one; s, number of spare components; R, total number of each component being operated.

Notice that this result is only for one type of unit. In a situation where unit i is actually made up of r_i components and there are M such units in operation, the total number of operating components is $R=\sum_{i=1}^{M} r_i$.

To address the problem of component redundancy, Sherbrooke [3] uses the hypergeometric distribution to properly distribute the total number of component failures, x, across all M operating units. The number of component failures, y_i, on each unit i is such that $\sum_{i-1}^{M} y_i = x$. The distribution of X is shown below:

$$\text{hyp}(y_1,y_2,y_3,\ldots,y_M|x)=\frac{\binom{r_1}{y_1}\binom{r_2}{y_2}\cdots\binom{r_M}{y_M}}{\binom{R}{x}} \tag{13.6}$$

where x, total number of component failures; r_i, number of components on each unit i; y_i, number of component failures on unit i such that $\sum_{i=1}^{M} y_i = x$.

In this chapter, redundancy can imply two things. First, within a unit i, redundancy implies that not all r_i components must operate for the unit to function properly. Thus, assume that for unit i to operate properly, only q_i of the r_i components must be operating. Secondly, system redundancy implies that in a system of M units, not all M units need to operate for the system to function properly. Therefore, assume that I units of the M must be operating as a minimum for system success.

To illustrate these concepts more clearly, let us look at a simple radio example (see Figure 13.1). This radio system is composed of transmitting and receiving units. Transmitter units for each radio are made up of three transmitter components ($r_1=3$). Receiver units are each made up of a single receiver component ($r_2=1$).

Based on the earlier discussions, the failure distribution of the units in this M radio system can be calculated. First, a failure distribution is needed for each unit type. Here we take the transmitter as one type ($j=1$) and the receiver as a second type ($j=2$). Since numbering of units is arbitrary, we assume the "first" fails. The failure distribution of unit type j, for the first 1, 2,...., k units up and the $k+1$, $k+2$, M units down, is calculated using:

$$U_j(k)=\sum_{x=0}^{\infty}P_j\{X=x\}\sum_{T}\text{hyp}(y_1,y_2,\ldots,y_M|x)\quad I\leq k\leq M \tag{13.7}$$

where j, the unit type number; T, the complete set of component failure combinations that result in the first k units of type j up and the rest down; I, the minimum number of radios that must be available.

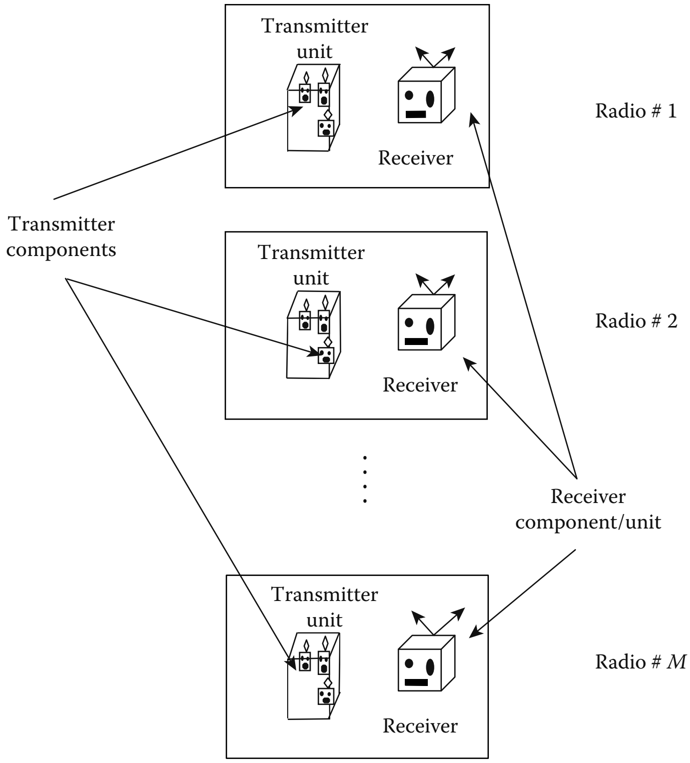


FIGURE 13.1 *M* radio operating system.

One advantage to this approach is that the set T of possible failure combinations, resulting from the first k units operating and the last $M-k$ being down, limits the number of hypergeometric calculations that must be performed. Also, since these calculations are not dependent on the failure or repair distributions, they can be calculated once and used for all components.

The results for each unit type are then combined into a radio system failure distribution. For the radios, there are two different unit types, a transmitter and receiver type. The probability that all M radios are up and none are down, $U(M)$, is given by:

$$U(M)=U_1(M)*U_2(M) \tag{13.8}$$

which is simply the probability of all the transmitter units being up times the probability of all the receiver units being up.

Similarly, $U(M-1)$ is the probability that the first $M-1$ radios are up and radio M is down.

$$U(M-1)=U_1(M-1)\,[U_2(M-1)+U_2(M)]+U_1(M)U_2(M-1) \tag{13.9}$$

This equation results from unit combinations (transmitter and receiver) that create the first $M-1$ radios up and the Mth radio down. For example, the first term in Equation 13.9 is created by the cases where either both unit types have the first $M-1$ units up or the transmitter unit type has the first $M-1$ up and the receiver unit type has all M up. Both of these combinations results in the first $M-1$ radios up and the Mth radio down. The equation for the first $M-2$ radios up and the $M-1$ and M radios down is more complex:

$$U(M-2)=U_1(M-2)[U_2(M-2)+2U_2(M-1)+U_2(M)]+2U_1(M-1)$$
$$[U_2(M-2)+U_2(M-1)]+U_1(M)U_2(M-2) \tag{13.10}$$

The logic for building this equation is the same as explained previously, that is, enumerate all combinations that produce the first $M-2$ radios up and the $M-1$ and Mth radios down. For example, the second term $(U_1(M-2) * 2U_2(M-1))$ is a result of the fact that with two radios down for transmitter units, there are two ways to select the radio that is also down for the receiver unit. This combinatorial equation building process continues until the minimum number of acceptable operating radios, I, is reached.

Based on the calculations for operating radio probabilities, an overall radio system availability can be calculated. Since $U(k)$ is the probability that the first k radios are operating while the remaining $M-k$ are down, the availability, $A(k)$, of at least k radios operating can be found by computing the number of ways in which k up radios can be selected from the M radios. This results in the following equation:

$$A(k)=100\left[U(M)+\binom{M}{M-1}U(M-1)+\ldots+\binom{M}{k}U(k)\right],\quad I\le k\le M \tag{13.11}$$

This is the only model, to date, that incorporates finite source arrivals, component redundancy, and multiple components into one availability model.

The preceding approach handles finite-population demand (i.e. failure rate decreases as failures increase) within a component-type, but fails to consider the impact the failures of each component type have on the other component-type failures. Therefore, some type of adjustment must be made to account for this inter-component interaction. We propose an iterative approach. An adjustment of the other component-type's failure rates based on the highest failing component-type's expected availability will reflect this interaction. The expected number of operating units, $E[O_j]$, is calculated for each component-type j using:

$$E[O_j]=\left[\sum_{k=0}^{M}\binom{M}{M-k}U_j(M-k)*(M-k)\right] \tag{13.12}$$

Then the component-type with the lowest number of expected operating units would be used to adjust the demand rates of the remaining component-types. The adjustment factor is the availability of this least available component-type. This availability is then multiplied by the arrival rate of all the other component-types to obtain a new utilization rate. This equation is shown below:

$$f_{j\,\text{new}}=f_{j\,\text{old}} * E[O_j]/M \tag{13.13}$$

This new utilization rate, $f_{j\,\text{new}}$, is used to recompute all the other component-type's $U_j(k)$. Then the system availability is recalculated, checking to see if the increase in $A(M)$ (i.e. $A_{\text{Current}}(M)-A_{\text{Last}}(M)$) is more than 0.001 (we find 0.1% to work well). If it is, then the second lowest $E[O_j]$ value is selected and new $f_{j\,\text{new}}$ values are computed for all but the lowest two $E[O_j]$ values. Then new availability's are computed. When the change in $A(M)$ is less than 0.001, stop the process and report the availability results.

This iterative procedure requires only a small fraction of the computational time necessary by other techniques (such as simulation) to consider all of the combinations of failure interactions by focusing on the most likely to fail components. If we had a system with a single class of unit, and no spares, we could exploit the k-out-of-n system structure (see Rushdi [7]), but our system is more complicated than that. Our iterative procedure dramatically reduces computational effort while including both redundancy and spares. Figure 13.2 and the algorithm given below summarize the method.

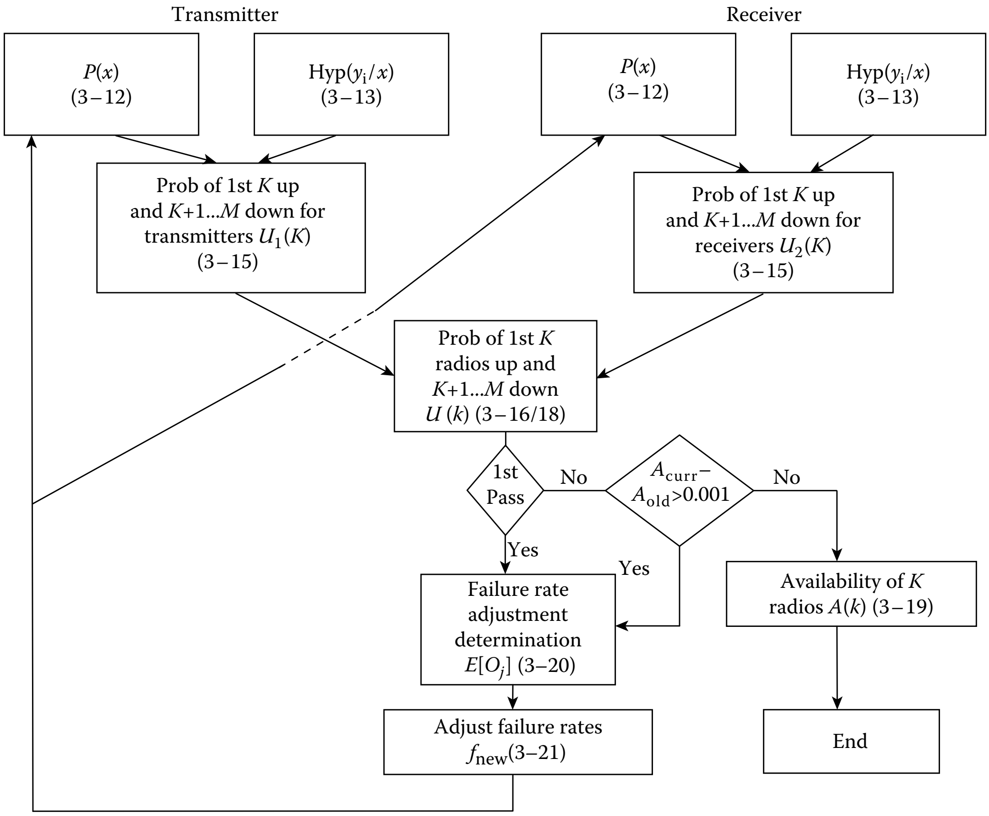


FIGURE 13.2 Flow of model calculations.

13.2.1 Algorithm

Step #1: Calculate the component failure distributions, $P(X)$, and the hypergeometric distribution of failures hyp(y/x) for each component (Equations 13.5 and 13.6).

Step #2: Calculate the probability that the first k components of each are up and the $k+1$, $k+2$,..., M are down, $U_j(k)$ (Equation 13.7).

Step #3: Calculate the probability that the first K units of the system are up and the $K+1$, $K+2$,..., M units are down, $U(k)$ (Equations 13.8 through 13.10).

Step #4: If first pass or $A_{current}(M)-A_{last}(M)>0.001$, go to step 5. Otherwise, go to step 8.

Step #5: Calculate the expected number of operating components of each component-type, $E[O_j]$ (Equation 13.12).

Step #6: Select the smallest $E[O_j]$ value in step 5 and calculate the new utilization rates, $f_{j\,new}$, for all the other component-types (Equation 13.13).

Step #7: Go to step 1 and repeat the steps for all components except the components whose $E[O_j]$ have already been computed.

Step #8: Calculate the availability, $A(k)$, for the system based on the desired number of operating pieces in the system (Equation 13.11).

We now have a method to overcome all the shortcomings of current logistics models for spares provisioning. The next section presents computational and validation results based on a Gulf War case study.

13.3 Case Study

Validation is always an important aspect of model development. In order to build a comparison database, a simulation model was developed using the SLAM II language [8] to create simulated data. The simulation model by Lewis [9] was used as a baseline for creating this simulated data. It had already been validated against operational data, so it provides an excellent starting point.

In addition, Headquarters Air Combat Command (HQ ACC) provided expert advice on data, modeling and validation issues. Aircraft in ACC are basically divided into two groups, fighters and "heavies". Fighters include aircraft such as the F-15, F-16, A-10 and F-111. Heavies include aircraft such as the B-52, B-1, E-3, and B-2. The test parameters were established through work with HQ ACC and are documented by Miyares [10]. Two aircraft from each group based on the recommendations of HQ ACC have been selected as typical of a deployed Gulf War unit. These aircraft are the B-52, E-3B, F-15, and F-16. For each of these aircraft types, three aircraft sizes were selected. Stock positions were set at three different levels: zero stock, fully authorized level, and the average fill percentage level based on March 1995 fill rates. Flight profiles were also provided by HQ ACC.

The results presented below are a typical sample of the 120 different cases we analyzed. In only three cases of 120 were the proposed model results outside of the 95% confidence intervals for the simulation model, and the largest 95% simulation confidence interval based on availability was less than 4% wide. In addition, hypothesis of means tests were performed on the two different redundancy case output differences and they both had high power values. This provides strong evidence that the new model is highly accurate. Furthermore, the run times for the analytical model required CPU seconds to execute, whereas the simulation model requires CPU hours to accomplish the same task.

13.4 An Optimizing Tool

Another capability missing from the current models is the ability to quantify any additional capability (or shortage) that might be available if the current availability goal is exceeded (or not met). For example, Dyna-METRIC will tell the user that he can fly his requested flying hours and maintain a 85% aircraft availability. But what if the availability goal is only 80%? Dyna-METRIC cannot tell you how many additional flying hours can be flown and yet still meet the availability goal. By adding a feedback mechanism to the proposed model and providing a user-input availability goal (Y), the proposed model can report the maximum operating hours per day per system, H, that can be achieved and still maintain the user availability goal (Y). This can be stated as:

$$\begin{aligned} \text{Maximize} \quad & H = \text{OHC} * \text{CPD} \\ \text{Subject to} \quad & A(M-1) \geq Y \text{ for } M \geq 3 \\ & A(M) \geq Y \text{ for } M < 3 \end{aligned}$$

where H, number of operating hour per system per day; Y, user-input availability goal; $A(M)/A(M-1)$, are defined in Equations 13.8 and 13.9; OHC, operating hours per cycle; CPD, cycles per day.

This problem is solved by adding the following step to the algorithm presented previously.

Step #9: If the availability of $M-1$ or more systems is greater than the availability goal, Y, (the availability of M systems is used for scenarios where $M<3$), increment the daily operating hours per system by the user provided step size, G, and go to step #1. If the availability goal, Y, is not exceeded, then report the availability and operating hours for the last acceptable operating hour/availability combination. If no such combination exists, simply report this operating hour program and the expected system availabilities, $A(k)$'s.

Of course, if the availability goal could not be met by the current stock quantities, the stopping rule is changed to decrement operating hours to find the maximum operating hours in order to meet the availability goal.

The following example illustrates how this capability might be used. In Tables 13.1 and 13.2, the B-52 three aircraft availability of two or more aircraft was greater than 90%. What if the availability goal was only 80%? How many additional flying hours per aircraft per day could someone achieve and still meet the availability goal? If this information is used in the optimization model, we discover that instead of 7.2 flying hours per aircraft, we can get 10.20 flying hours per aircraft while achieving a 81.44% availability. This process could also be used to study the relationship between different stock levels and flying hours and their impact on aircraft availability using the optimization model.

Figure 13.3 shows a complete response surface for the relationship between different stock levels, flying hours, and aircraft availability for the B-52 three aircraft case discussed above. Notice that it is unimodal (a property we observed in all cases), so that the optimization model yields the global solution.

Here we have a complete view of the general relationship between stock levels, flying hours, and availability which we could not create with current modeling capability. This graph was generated by running the model with 1% availability goals and a step size of 0.5 flying hours. The stock levels were decremented from 100% by randomly selecting the missing items for each additional loss in available stock until zero was reached. This graph allows the general assessment of any combination of flying hour and stock level decisions. However, it is much more computational demanding than a single search using the optimization model.

TABLE 13.1 No Redundancy Comparison of Results, Full Stock Case

				Simulation 95% CI		
	# of A/C	Research Model	Model Error	Lower	Mean	Upper
E3B	5	19.40	0.06	18.14	19.34	20.54
		37.50	−0.06	36.49	37.56	38.62
	3	48.28	0.07	46.91	48.21	49.51
		39.71	0.06	38.77	39.65	40.53
	1	82.38	−0.43	81.84	82.81	83.79
F15	5	24.33	0.99	22.35	23.34	24.33
		39.67	0.22	38.66	39.45	40.24
	3	53.90	1.18	50.99	52.72	54.45
		36.94	−1.09	36.52	38.03	39.53
	1	86.89	−0.23	85.70	87.12	88.54
B52	5	25.78	−0.33	25.34	26.11	26.89
		37.97*	−0.73	38.20	38.70	39.20
	3	54.34	0.07	53.29	54.27	55.25
		35.67	−0.60	35.65	36.27	36.89
	1	82.38	−0.09	81.45	82.47	83.49
F16	5	87.97	−0.42	87.49	88.39	89.29
		11.41	0.48	10.11	10.93	11.75
	3	94.76	−0.46	94.50	95.22	95.93
		5.142	0.459	3.997	4.683	5.368
	1	97.73	0.27	96.89	97.46	98.04

*Outside of the simulation 95% confidence interval.

TABLE 13.2 Redundancy Case Comparison of Results, Partial Stock Case

				Simulation 95% CI		
	# of A/C	Research Model	Model Error	Lower	Mean	Upper
E3B	5	17.67	0.07	16.81	17.60	18.38
		36.55	0.34	35.49	36.21	36.93
	3	45.47	0.26	43.58	45.21	46.85
		40.93	0.44	39.19	40.49	41.79
	1	77.63	−0.04	76.3	77.67	79.04
F15	5	19.36	0.54	17.96	18.82	19.67
		37.58	0.1	36.67	37.48	38.28
	3	46.46	0.71	43.96	45.75	47.54
		40.55	−0.13	39.21	40.68	42.16
	1	78.46	1.01	76.06	77.45	78.94
B52	5	16.31	0.46	14.96	15.85	16.74
		35.19	0.26	33.97	34.93	35.89
	3	56.18	−0.2	55.05	56.38	57.7
		34.61	−0.35	33.94	34.96	35.98
	1	82.72	−0.3	82.03	83.02	84.01
F16	5	76.52	0.31	74.64	76.21	77.77
		20.88	0.08	19.46	20.8	22.15
	3	85.56	−0.48	85.05	86.04	87.04
		13.69	0.42	12.33	13.27	14.21
	1	95.49	0	94.69	95.49	96.29

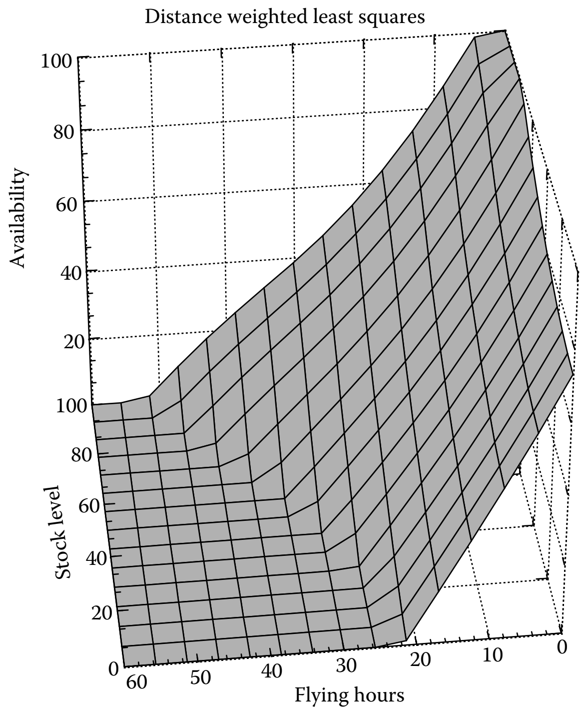


FIGURE 13.3 Relationship of flying hours and stock level to availability.

13.5 Summary

Logistics support has always been a key element of combat effectiveness as well as an important element in commercial aircraft systems. The Air Force is interested in developing tools and methods to assess the impact of logistics on combat capability. The current method used by the Air Force to assess aircraft spare's support was developed for use with a large number of aircraft. With the high cost of new aircraft, fewer aircraft are fielded and, therefore, some of the critical assumptions made in the current model are unsatisfactory. This includes the assumption of an infinite calling population of demands and the distribution of backorders by sampling with replacement. This research develops an original method to assess the system availability of a small number of vehicles or machines which overcomes the flaws of the existing model. The new method also allows the user to include component redundancy and component spares in the system.

In order to test the research method, a FORTRAN program was written to implement the approach. Using actual US Air Force failure, repair, and scenario data for the F-16, F-15, E-3B, and B-52H, the results were compared to simulation results. The research model was tested under two different cases: component redundancy and no component redundancy. In both cases, the research model performed very closely to the simulated data. This provides strong evidence that the research method performs well under a wide range of operating conditions. The research model has an efficiency advantage over the simulation model which include speed of processing (performance improvement of several orders of magnitude) and, thus, the ability to handle more components in the system for a fixed response time.

Finally, an additional capability was added to the research method that does not exist in the current model or in the simulation model. An optimization technique is provided that allows the user to optimize operating hours given a system availability goal or target. This new ability allows the user to estimate the additional capability available (i.e. additional operating hours), or shortage, if he is currently not matching his availability goal.

References

1. Isaacson, K., Boren, P., Tsai, C., and Pyles, R. 1988. Dyna-METRIC Version 4: Modeling Worldwide Logistics Support of Aircraft Components. The RAND Corp., R-3389-AF.
2. Palm, C. 1938. Analysis of the Erlang traffic formulae for busy-signal arrangements. *Ericsson Technics,* 4, 39–58.
3. Sherbrooke, C. 1992. *Optimal Inventory Modeling of Systems: Multi-Echelon Techniques*. John Wiley & Sons, New York, p. 274.
4. Faughaber, K. 1993. DMAS Small PAA Utility. HQ USAF/LGSS, Washington DC.
5. Pyles, R. and Tripp, R. 1983. Dyna-METRIC: A valuable management tool. *Air Force Journal of Logistics,* Spring, 18–21, 24.
6. Kaplan, A. 1989. Incorporating redundancy considerations into stockage policy models. *Naval Research Logistics,* 36, 625–638.
7. Rushdi, A.M. 1986. Utilization of symmetric switching functions in the computation of k-out-of-n system reliability. *Microelectronics and Reliability,* 26(5), 973–987.
8. Pritsker, A. 1986. *Introduction to Simulation and SLAM II*. John Wiley & Sons, New York.
9. Lewis, T. 1987. Unit Level WRSK Assessment and Sortie Generation Simulation Model. MS Thesis, Air Force Institute of Technology, Wright Patterson AFB, OH.
10. Miyares, Capt. P., 1995. Validation of Aircraft Combat Capability Model—Memorandum. HQ ACC/LGSW, Langley AFB, VA, 24 March 1995.

14

High Velocity Maintenance: The Role of Industrial Engineering in USAF Maintenance, Repair, and Overhaul

Doug Keene
Robins Air Force Base

Frank Dement
Robins Air Force Base

Gary O'Neill
Robins Air Force Base

The Air Force is recognized as a world leader in pushing the limits of technology to overcome constraints to mission success. The weapon system development efforts resulting in the F-117, B-2, F-22 and Global Hawk UAV are clear examples of innovative advanced technologies leading to game-changing effects in the battlespace. Despite the USAF's cutting edge technological supremacy, the maintenance, repair, and overhaul (MRO) of USAF weapon systems has experienced only limited advancement through the application of the full spectrum of industrial engineering (IE) methods. While USAF MRO operations have made excellent progress through implementing Lean and Air Force Smart Operations for the 21st Century (AFSO21) methodologies, these functions and facilities still largely operate as "job shops". In this environment, the craftsman/mechanic accomplishes all work site preparation; gathers parts, tools and equipment for the day's tasks; acquires all Technical Order (TO) manuals, drawings, and documentation required for each task; and performs a variety of other activities that limit overall productivity.

Under AFSO21, USAF MRO operations have achieved incremental improvement in developing mechanic-centric processes that maximize value-added work and subsequently decrease downtime for the aircraft. With the application of the full range of IE methodologies, there is the potential for significantly larger improvements in efficiencies, leading to lower cost and increased aircraft availability. Achieving significant improvement in MRO to date has been elusive, given the prevailing mindset which perceives aircraft heavy maintenance and repair as so highly variable that detailed planning and scheduling is virtually impossible. With the increased emphasis in today's USAF on waste reduction and standard work, combined with the progress already achieved by MRO operations through Lean and AFSO21, the Air Force is well postured to take aggressive steps in changing this prevailing mindset.

Commercial airlines have long recognized that incoming revenue stops when their aircraft are on the ground for maintenance. Given the highly competitive nature of the business, commercial airlines are particularly sensitive to lost opportunities for revenue generation. Consequently, the airlines have devoted significant time and energy in applying IE methods, such as standard work and parallel efforts, to reduce aircraft downtime. Commercial airlines' organic and inorganic MRO operations have eliminated sources of waste in order to ensure mechanics are adding value at the maximum possible rate. This "burn rate", or number of touch labor hours per day, has been acknowledged by commercial MRO organizations as the critical factor for reduction of aircraft downtime. To maximize burn rate, commercial MRO activities optimize work procedures while the aircraft is in the repair dock through continuous, extensive and detailed planning and feedback. Aptly described as choreographed, the level of detail to which they plan personnel and parts movements enables commercial MRO to accomplish work at a rate up to eight times faster than comparable USAF Programmed Depot Maintenance activities. Regularly achieving this high level of MRO burn rate has resulted in the airlines routinely maintaining aircraft availability rates well above 90%, compared to the approximately 60% availability rate the USAF typically maintains.

The Air Force has begun an initiative called High Velocity Maintenance (HVM) to use IE methods and Lean principles to dramatically increase aircraft availability, in emulation of commercial aircraft MRO. Two primary changes in the current state of practice must occur to enable HVM: (1) more complete understanding of asset condition prior to induction in order to effectively and timely plan and resource maintenance and repair actions; (2) efficient, standard work processes to accomplish maintenance and repair. The latter will be achieved primarily through the application of IE methods to enable mechanic-centric support. This support includes engineered standards; work stations designed for efficiency; ergonomic tools and setups; more effective technical data; automated data collection and retrieval; ready access to expertise to resolve technical issues; and complete, specialized task kits to affect repair without leaving the task site. HVM is expected to increase the velocity of aircraft through the MRO operations achieving burn rates approaching 500 hours per day, four times the present rate. On the C-130 fleet alone this could result in up to 55 aircraft being returned to the field and could increase aircraft availability by 14%. Just as significant is a potential reduction in cost of PDM by $387M over six years.

MRO of a complex weapon system involves a large number of inter-related tasks. Accumulated damage to the airframe, the inspection for and repair of being the primary motivation for MRO, manifests itself as cracks, corrosion, fretting, wear, or other structural anomalies. Factors which greatly influence damage accumulation include not only the operational usage spectrum and environmental conditions experienced by the individual aircraft, but also aircraft age. Given these driving factors, conventional wisdom would indicate that MRO has a high degree of variability. When the MRO work packages for aircraft are examined, the individual repair actions are often quite variable. However, much of the work package is, in fact, quite repetitive, such as inspection, disassembly, reassembly, depaint/paint, and functional testing. While IE techniques are applicable to the entire process, they are particularly valuable in those repetitive processes which lend themselves to efficiency improvements through standard processes. Even in the highly variable portions of the MRO work package, there are ample opportunities to dramatically improve task efficiency by enabling the mechanic to stay on task versus the current paradigm where the mechanic is saddled with limited value-added preparation work.

The initiative to implement HVM has identified a key opportunity for USAF MROs: effective utilization of IEs in process improvement. In USAF MRO operations, credentialed IEs have historically not been involved in the setting of labor standards, the planning of the work package, or establishing of labor requirements. This, in itself, tends to separate the IE from the core business decisions which lead to the discovery of inefficiencies in the process. Consequently, USAF IEs have traditionally migrated to those activities that would more readily align with duties prescribed for credentialed mechanical engineers. There is a dire need to incorporate IEs into USAF MRO process improvement and more effectively leverage their skills to enhance the progress that can be made under AFSO21. One aspect of the HVM initiative is giving IEs direct responsibility for these aspects of process improvement.

In order to achieve the efficiencies necessary to dramatically improve aircraft availability and lower cost, the USAF must aggressively apply IE principles and more effectively employ IEs in the planning and production improvement process. It is imperative that IEs be involved in developing engineered work standards and work stations that allow high efficiency task completion. IEs, using operations research methods and statistical analysis, must apply algorithms to maximize labor application. With their toolkit of AFSO21 techniques, USAF IEs will need to engage with the workforce to prepare detailed work standards and methods that can be scripted and supported with task kits containing all the elements needed for task completion. It is also crucial that IEs actively and continuously assist production managers to make maximum use of parallel operations when applicable. Lastly, USAF IEs will need to design facilities and work centers that exploit robotics and machine-assisted operations to improve task velocity and quality. In short, robust application of IE methods, and greater utilization of the IE workforce, will have significant positive effects on USAF MRO velocity, cost, and quality, and ultimately aircraft availability. Our customers deserve nothing less.

15

Beyond Authorized versus Assigned: Aircraft Maintenance Personnel Capacity*

Jeremy A. Howe
Air Force Logistics Management Agency

Benjamin A. Thoele
Air Force Logistics Management Agency

Scotty A. Pendley
Air Force Logistics Management Agency

Anthony F. Antoline
Air Force Logistics Management Agency

Roger D. Golden
Air Force Logistics Management Agency

15.1 Introduction

Most would agree that aircraft maintenance has been and continues to be a challenging, complex task involving a delicate balance of resources to include personnel, equipment, and facilities. This balancing act occurs in a very hectic environment. The Air Force flies 430 sorties per day in support of Operation Iraqi Freedom and Enduring Freedom. A mobility aircraft takes off somewhere in world approximately every 90 seconds [1]. As the demand for aircraft continues to grow, the number of airmen who support these aircraft is declining. "Since 2001 the active duty Air Force has reduced its end-strength by almost 6% but our deployments have increased by at least 30%, primarily in support of the Global War on Terror" [2]. This reduction in personnel is part of the Air Force's process of drawing down the total force by approximately 40,000 people, with many of these cuts in aircraft maintenance career fields. Also adding to the growing maintenance workload is an aircraft fleet which now average almost 24 years old, with the average age still increasing [3].

When it comes to aircraft maintenance, the Air Force depends on metrics to know whether or not we are measuring up to standards. Several metrics exist which attempt to measure the success or failure of our maintenances' efforts. One of the most recognized metrics is the total not mission capable maintenance (TNMCM) rate. Air Force Instruction 21-101 describes TNMCM as "perhaps the most

* This material was originally published in the *Air Force Journal of Logistics,* XXXI (4), Winter 2008, 14–24. Reprinted with permission.

common and useful metric for determining if maintenance is being performed quickly and accurately" [4]. Although a lagging type indicator, it is one of several key metrics followed closely at multiple levels of the Air Force. Over the last several years, the TNMCM rate for many aircraft gradually increased. This fact was highlighted during a 2006 quarterly Chief of Staff of the Air Force Health of the Fleet review. Follow-on discussions ultimately resulted in the Air Force Material Command Director of Logistics (AFMC/A4) requesting the Air Force Logistics Management Agency (AFLMA) to conduct an analysis of TNMCM performance with the C-5 Galaxy aircraft as the focus. AFLMA conducted two studies in support of this request.

15.2 Background

The C-5 TNMCM Study II (AFLMA project number LM200625500) included five objectives. One of those objectives was to determine root causes of increasing TNMCM rates for the C-5 fleet. An extensive, repeatable methodology was developed and utilized to scope an original list of 184 factors down to two potential root causes to analyze in-depth for that particular study. These two factors were aligning maintenance capacity with demand, and the logistics departure reliability versus TNMCM paradigm. To address the root cause factor of aligning maintenance capacity with demand, a method of determining available maintenance capacity was needed. To meet this objective, a new factor designated as net effective personnel (NEP) was developed. NEP articulates available maintenance capacity in a more detailed manner that goes beyond the traditional authorized versus assigned personnel viewpoint. The remainder of this article describes the need for NEP and how the NEP calculations were developed during the C-5 TNMCM Study II. The NEP calculations were ultimately used in conjunction with historical demand to propose base-level maintenance capacity realignments resulting in projected improvements in the C-5 TNMCM rate.

15.3 Personnel as a Constraint

The analytical methodology applied to the C-5 maintenance system determined that personnel availability was an important factor to consider. This idea is not new; indeed, the force-shaping measures underway in the Air Force have brought the reality of constrained personnel resources to the fore front of every airman's mind. Without exception, maintenance group leadership (MXG) at each base visited during the C-5 TNMCM Study II considered personnel to be one of the leading constraints in reducing not mission capable maintenance hours. The study team heard the phrase "we need more people" from nearly every shop visited:

> "The biggest problem for the maintainers here is a shortage of people" [5].
>
> "With more people we could get a higher MC [mission capable]. We're currently just scrambling to meet the flying schedule" [6].
>
> "Hard-broke tails and tails in ISO [isochronal inspection] get less priority than the flyers. We run out of people—we physically run out" [7].

The Air Force defines total maintenance requirements (authorizations) on the basis of the logistics composite model (LCOM) and current manpower standards. LCOM is a stochastic, discrete-event simulation which relies on probabilities and random number generators to model scenarios in a maintenance unit and estimate optimal manpower levels through an iterative process. The LCOM was created in the late 1960s through a joint effort of RAND and the Air Force Logistics Command. Through intended to examine the interaction of multiple logistics resource factors, LCOM's most important use became establishing maintenance manpower requirements. LCOM's utility lies in defining appropriate production levels, but it does not differentiate experience [8]. Once these requirements are defined, the manpower community divides these requirements among the various skill levels as part of the programming process. Overall, the manpower office is charged with determining the number of slots, or spaces, for each skill level needed to meet the units' tasks. The personnel side then finds the right *faces*, or people, to fill the spaces.

One measure historically used to quantify personnel availability is the ratio between authorized and assigned personnel. While this ratio is an indicator of maintenance capacity, it provides only a limited amount of information. Authorized versus assigned ratios do not take into account the abilities and skill levels of the maintenance personnel, nor does it factor in the availability of the personnel on a day-to-day basis. These issues were addressed in the C-5 TNMCM Study II by quantifying "we need more people" beyond the traditional metric of authorized versus assigned personnel. This capacity quantification was done as part of the larger effort of aligning capacity with demand. The process of capacity planning generally follows three steps:

- Determine available capacity over a given time period.
- Determine the required capacity to support the workload (demand) over the same time period.
- Align the capacity with the demand [9].

The following describes how the study team pursued step 1, determining available capacity over a given time period, using data from the 436 MXG at Dover Air Force Base (AFB) and characterizing the results in terms of what the study team denoted as NEP.

15.4 Determining Available Capacity

When personnel availability and capacity are discussed at the organizational level, typically the phrase *authorized versus assigned* personnel is used. However, are all people assigned to maintenance organizations—namely, an aircraft maintenance squadron (AMXS) or a maintenance squadron (MXS) —viable resources in the repair process? Most maintainers will answer no. While it is true that all assigned personnel serve a defined and important purpose, not everyone in these organizations is a totally viable resource to be applied against maintenance demand. This impacts maintenance repair time and aircraft availability.

TNMCM time begins and ends when a production superintendent advises the maintenance operations center to change the status of an aircraft. The length of that time interval is determined by several things. One factor is the speed of technicians executing the repair, which includes diagnosis, corrective action, and testing (illustrated in Figure 15.1) the repair node of Hecht's *restore-to-service* process model.

As illustrated by the Hecht process model, there are other important components required to return an aircraft to service, but the pool of manpower resources required to support the repair node is critically linked to TNMCM time. Within a mobility aircraft maintenance organization, this pool represents hands-on 2AXXX technicians whose primary duty is performing aircraft maintenance. Specifically, the study team defined the technician resource pool as follows:

> **Technicians**: the collective pool of airmen having a 2AXXX AFSC, that are 3-level or 5-level maintainers, or nonmanager 7-level maintainers whose primary duty is the hands-on maintenance of aircraft and aircraft components.

The distinction of nonmanager 7-levels generally reflects 7-levels in the grades of E-5 and E-6. In active duty units, 7-levels in the grade of E-7 do not typically perform hands-on aircraft maintenance, but are instead directors of resources and processes—they are managers [10]. This is in stark contrast to Air National Guard units, where 2AXXX personnel in the senior noncommissioned officer ranks routinely perform *wrench-turning*, hands-on maintenance [11]. For the research detailed in the C-5 TNMCM Study II, personnel analysis centered on data from the 436 MXG at Dover AFB and utilized the study team's definition of technicians.

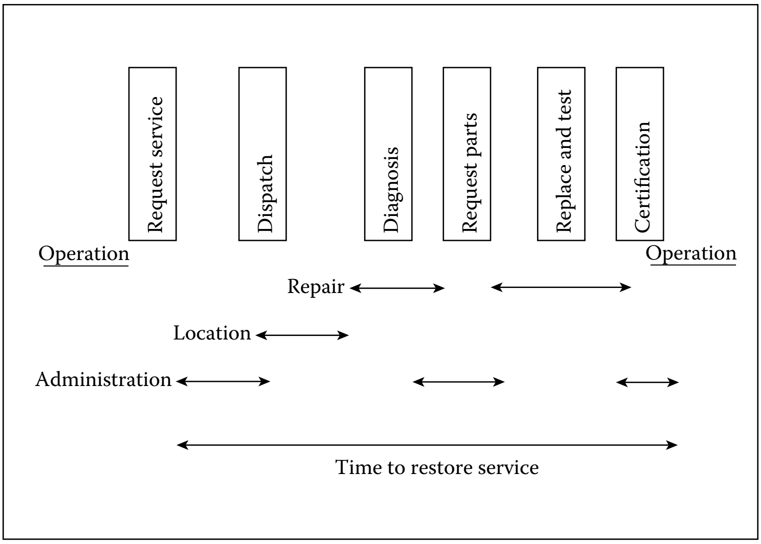


FIGURE 15.1 Time to restore service process model. (From H. Hecht. 2004. *Systems Reliability and Failure Prevention*. Norwood, MA: Artech House.)

15.5 Net Effective Personnel

Authorized versus assigned personnel figures usually quantify the entire unit. With the definition of technicians in mind, it is important to consider three additional factors that introduce variability into the personnel resource pool. These factors are:

- Skill-level productivity.
- Ancillary and computer-based training (CBT).
- Availability.

The study team examined the influence of these three factors, as well as their impact on the viable resource pool for the 436 MXG. This collective impact yielded a new resource pool representing a depiction of *effective* capacity rather than just the authorized versus assigned ratio. Again, this new resource pool is denoted as net effective personnel, or NEP.

15.6 Factor 1: Skill-level Productivity

In order to accurately examine the quantitative adequacy of a resource, as well as how a resource has historically been used to meet demand, there must be parity among individual resource units. Consider the previous definition of technicians. If one were to select two people at random, would they be equally capable resources? Not necessarily, if one was a 3-level trainee and the other was a 5 or 7-level resource. In order to collectively examine people in terms of comparable resources, and to account for the skill-level variability in typical aircraft maintenance organizations, productivity factors were applied to the resource pool.

As part of this research effort, the study team utilized its strategic partnership with RAND Project Air Force. Through personal interviews with RAND personnel and review of recently published RAND research, the study team learned that RAND had explored the productivity of trainees and trainers in aircraft maintenance units. Trainees were defined as 3-levels, who are not as productive as 5- and 7-levels. Additionally, some 5- and 7-levels were not as productive as others because they spend time training and instructing 3-level personnel [12]. In terms of specific productivity based on RAND research, 3-levels were estimated to be 40% productive, 5-level trainers and nonmanager 7-level trainers were estimated to be 85% productive, and 5-levels and nonmanager 7-levels were 100% productive if they were unencumbered with training responsibilities [13]. For the purpose of this analysis, the number of trainers was considered to be equal to the number of 3-levels assigned—a one-to-one ratio. The productivity factors for the viable resource pool are summarized in Table 15.1.

These productivity factors also are similar to results from additional RAND research at Travis AFB published in 2002 [14]. Considering the productivity factors from Table 15.1, the net effect of these productivity factors alone was a reduction of the 436 AMXS viable resource pool by an average 5.68% [15].

15.7 Factor 2: Ancillary Training and Computer-based Training

In recent times the impact of ancillary training and CBT has been such an important issue for Air Force senior leaders, that it was the sole topic of the airman's Roll Call of 9 February 2007 [16]. This document indicated that some active duty airmen spend disproportionate amounts of time on ancillary training, which detracts from their ability to perform official duties. Moreover, the document suggested that some ancillary training may no longer be relevant [17]. In the context of the viable pool of aircraft maintenance technicians, this would mean that, some of the time, personnel resources may be on duty but unavailable to perform hands-on maintenance due to an ancillary training requirement.

A consensus majority of personnel interviewed during the study team's site visits echoed these concerns, describing an *insidious growth* of new training requirements in recent years [18]. An additional concern voiced by interviewees pertained to computer resources. Interviewees described a situation where office workers have ready access to a personal computer (PC) but dozens of maintenance technicians often share only a handful of communal PCs. Consequently, their ability to complete computer-based ancillary training is constrained. One unit training manager explained that in the past, a group training briefing would be conducted for an entire work center, fulfilling each individual's training requirement simultaneously [19]. Today, an online course issues the required certificate of completion for only one individual, thereby necessitating that each airman conduct the training individually. The net result is more time away from primary duties (for example, repairing aircraft). In order to assess the influence of ancillary training and CBT on the technician resource pool, the study team quantified the average daily impact.

A list of various ancillary and computer-based training items that are applicable to the relevant pool of aircraft maintenance personnel was collected from three data sources:

- The USAF Education and Training Course Announcement (ETCA) website [20]
- The unit training monitor at the AFLMA
- The unit training monitor for the 105 MXG at Stewart Air National Guard Base (ANGB)

TABLE 15.1 Productivity Factors

Technician Category	Productivity Factor
Nonmanager 7-levels	100%
Nonmanager 7-levels trainers	85%
5-levels	100%
5-level trainers	85%
3-levels	40%

Source: J. G. Drew, K. F. Lynch, J. Masters, R. S. Tripp, and C. R. Roll, Jr. 2008. *Maintenance Options for Meeting Alternative Active Associate Unit Flying Requirements.* Santa Monica, CA: RAND Corporation, MG-611-AF, 2008.

The training was categorized by data source, course number (if applicable), and course name. Training was also categorized as follows.

- Mandatory for all personnel, such as law of armed conflict training
- Voluntary or job-specific, such as hazardous material management training

Also, requirements were identified by the recurrence frequency (one-time, annual, or semiannual). Some requirements are aligned with the 15-month aerospace expeditionary force cycle; this would equate to a yearly recurrence frequency of 0.8 (12/15). Finally, training was categorized by the duration in hours for each requirement as identified by the data sources.

Most training courses only take up a portion of the duty day. The average duration for courses considered was 2.8 hours, with many listed at 1 hour or less. In situations like these, a manager would still view the individual as *available* for the duty day [21]. Therefore, the study team examined the impact of CBT and ancillary training as a separate factor and not as a part of the availability factor (Factor 3). Final calculations resulted in the following totals:

- Hours of mandatory one-time training (denoted M_o), 101.5 hours
- Hours of mandatory annually recurring training (M_a), 67.2 hours
- Voluntary or job-specific one-time training (VJS_o), 85.8 hours
- Voluntary or job-specific annually recurring training (VJS_a), 10.3 hours

In order to quantify the daily impact of these training items, the study team made the following assumptions:

- An eight-hour workday
- 220 workdays in a calendar year. (5 days per week × 52 weeks per year) = 260; 260−(30 days annual leave)−(10 federal holidays [22]) = workdays
- 3-levels required all of the mandatory, one-time training
- 5-levels and 7-levels required only the annually recurring portion of the mandatory training
- As an average, all 3-levels required 10% of the voluntary or job-specific, one-time training
- As an average, all 5-levels and 7-levels required 10% of the voluntary or job-specific, one-time, annually recurring training
- As an average, all training durations would be increased 20% to account for travel, setup, and preparation. (It should be noted that the study team performed sensitivity analysis on the last three assumptions, see Table 15.3.)

When employing the above assumptions, the figures in Table 15.2 were calculated to be best estimates of the time impact of ancillary training and CBT. The best estimates for CBT and ancillary training requirements account for 7.51 and 5.24% of the workday for 3-, 5-, and 7-levels, respectively. The complementary effectiveness rates for this factor are expressed at 0.9249 (1-0.0751) for 3-levels and 0.9476 (1-0.0524) for 5 and 7-level. These rates are listed as the ancillary and CBT factors for 3-, 7-, and 5-levels, respectively in Table 15.6.

Table 15.3 illustrates how these rates change when the percentages of voluntary and job-specific training (*V/JST*) or the percentage of travel and setup buffer are varied. The matrices in Table 15.3 illustrate the results of sensitivity analysis of various CBT and ancillary training factors that would result for combinations of

TABLE 15.2 Best Estimate of CBT and Ancillary Training Time Requirements

Technician	Hours per Year	Hours per Workday	Percentage of 8-hour Workday	Minutes per Workday
3-level	132.10	0.60	7.51%	36.03
Formula	$1.2(M_o+(0.1VJS_o))$	(hours/year)/220	(hours/workday/8) * 100	(hour/workday) * 60
5- / 7-level	92.17	0.42	5.24%	25.1
Formula	$1.2(M_q+(0.1(VJS_o+VJS_a))$	(hours/year)/220	(hours/workday/8) * 100	(Hrs/workday) * 60

TABLE 15.3 CBT and Ancillary Training Factor Sensitivity Analysis

3-Levels % Travel/Setup Multiplier						
%*V/JST*	1	1.05	1.1	1.15	1.2	1.25
0.00	0.942	0.939	0.937	0.934	0.931	0.928
0.05	0.940	0.937	0.934	0.931	0.928	0.925
0.10	0.937	0.934	0.931	0.928	0.925	0.922
0.15	0.935	0.932	0.929	0.925	0.922	0.919
0.20	0.933	0.929	0.926	0.922	0.919	0.916
0.25	0.930	0.927	0.923	0.920	0.916	0.913
5- and 7 Levels % Travel/Setup Multiplier						
%*V/JST*	1	1.05	1.1	1.15	1.2	1.25
0.00	0.962	0.960	0.958	0.956	0.954	0.952
0.05	0.959	0.957	0.955	0.953	0.951	0.949
0.10	0.956	0.954	0.952	0.950	0.948	0.945
0.15	0.954	0.951	0.949	0.947	0.944	0.942
0.20	0.951	0.948	0.946	0.944	0.941	0.939
0.25	0.948	0.946	0.943	0.940	0.938	0.935
Descriptive Statistics						
	Mean	Min	Max	Range		
3-Level	0.928	0.913	0.942	0.030		
5- and 7-Level	0.949	0.935	0.962	0.027		

voluntary or job-specific training, or travel and setup buffer ranging from 0 to 25%. The range of all calculated factors is approximately 3% for both technician categories. Note that the CBT and ancillary training factors chosen utilizing the study team's assumptions are boxed and shaded. For both 3-, 5-, and 7-levels, the calculated training factors fall very near the mean developed in the sensitivity analysis. Some values shown in Table 15.3 are the result of rounding. For the 436 MXG at Dover AFB, the net effect of these CBT and ancillary training factors alone was a reduction of the viable resource pool by an average of 1.58% [23].

15.8 Factor 3: Availability

Manpower resources must be present to be viable, and on any given day, aircraft maintenance organizations lose manpower resources due to nonavailability. Examples include temporary duty (TDY) assignments, sick days, and other details. To illustrate, Figure 15.2 depicts the actual availability of 436 AMXS airframe and powerplant general (APG) technicians on day shift for Thursday, April 12 2007. For this work center, on this particular day and shift, roughly 65% of assigned technicians were not available for the various reasons listed.

Much like aircraft maintenance, some events that take people away from the available pool are scheduled and known well in advance, while others are unexpected, such as illnesses and family emergencies.

Although scheduled and unscheduled events both have an impact, scheduled events are anticipated and can be planned for. Adjustments can be made and resources can be shifted. Consequently, resource managers want to monitor and manage scheduled personnel nonavailability to the greatest extent possible. In order to assess the impact of this factor on the resource pool, the study team monitored the personnel availability of the 436 AMXS at Dover AFB from 1 March through 30 April 2007 via nine weekly snapshots. 436 AMXS supervision tracks manpower via a spreadsheet tool that identified the availability status of each assigned 3-level, 5-level, and nonmanager 7-level in their hands-on maintenance resource

		3-level	5-level	7-level	Total	% of total
	Assigned	32	28	22	82	100%
Reason unavailable	Temporary duty		6	4	10	12%
	Qualification and training Program	9			9	11%
	Detail	2	3	2	7	9%
	Leave	2	3	2	7	9%
	Scheduled off day	2	1	2	5	6%
	Medical profile		2	1	3	4%
	Part-day appointment	1	1	1	3	4%
	Full-day appointment			2	2	2%
	Compensatory off day			1	1	1%
	Flying crew chief mission		1		1	1%
	Out processing		1		1	1%
	Permanent change of assignment		1		1	1%
	Field training detachment course		1		1	1%
	First term airmen's center	1			1	1%
	Bay orderly	1			1	1%
	Available	14	8	7	29	35%

FIGURE 15.2 436 AMXS APG day shift personnel availability snapshot. (From Data for 12 April 2007. 436 AMXS/MXAA, Dover AFB.)

pool. For AMXS, this represents technicians from six different shops, identified with the corresponding Air Force specialty codes (AFSC) as follows:

- Airframe and Powerplant General (APG)–2A5X1C, 2A5X1J
- Communication and Navigation (C/N)–2A5X3A
- Electro/Environmental System (ELEN)–2A6X6
- Guidance and Control (G/C) (G/C is alternatively known as Automatic Flight Control and Instruments)–2A5X3B
- Hydraulics (HYD)–2A6X5
- Engines (JETS)–2A6X1C, 2A6X1A

The AMXS snapshot spreadsheet is updated (but overwritten) continually as status changes occur [24]. By monitoring changes in these snapshots, the study team was able to examine not only the impact of personnel nonavailability in aggregate, but also the degree to which the discovery and documentation of events altered the size of the capacity pool. Using the Dover AMXS snapshots, the study team calculated the number of available technicians in the aircraft maintenance resource pool.

The study team monitored the actual availability figures for the 436 AMXS over the nine-week period of March and April 2007, for a total of $n=61$ daily observations. Across all shifts, the total number of personnel assigned to the AMXS personnel resource pool was 411 for the month of March, and 412 for the month of April. Actual availability figures, however, were much lower. Table 15.4 summarizes the descriptive statistics of this analysis.

The upper row of Table 15.4 statistics reflects the actual number of technicians available, while the bottom row reflects that number as a percentage relative to the total number of technicians assigned. For example, in the month of March, the maximum number of available technicians observed was 202, or 49% (202 of 411) of the total assigned. The mean availability of March was 36%. These figures take into consideration that some of the nonavailable personnel may be performing duties elsewhere for the Air

TABLE 15.4 436 AMXS Availability Descriptive Statistics

411 Assigned	March 2007				April 2007				March–April 2007			
	Min	Max	Mean	Range	Min	Max	Mean	Range	Min	Max	Mean	Range
Available	100	202	147	102	104	163	137	59	100	202	142	102
% of Assigned	24%	49%	36%	25%	25%	40%	33%	14%	24%	49%	35%	25%

Force such as flying crew chief missions or other TDY assignments. Therefore, they would not be viable assets for the aircraft maintenance resource pool at Dover AFB. The net effect of this nonavailability factor was a reduction of the AMXS home station viable resource pool by an average of 65.39%. This is reflected as the 35% mean highlighted for March–April 2007.

As discussed previously with Factor 1 and 2, the productivity of available technicians is reduced to skill-level training needs, as well as ancillary and CBT training requirements. The study team applied productivity factors from Table 15.1 and CBT and ancillary training factors from Table 15.2 to the observed number of available technicians in AMXS. These calculations quantified the final pool of viable personnel resources, which is denoted as NEP. Because of daily variations in the number of 3-, 5-, and 7-skill level technicians available, the factors were applied to each daily observation. In performing these calculations, the study team developed a representation of the effective personnel resource pool. Specifically, the NEP figures account for the realities of availability and productivity, and allow the resource pool to be viewed objectively, unconstrained by concerns such as skill-level differences. The value of such a resource picture is that it provides a suitable mechanism for comparing maintenance capacity (NEP resource pool) with maintenance demand. The summary descriptive statistics for the 436 AMXS NEP are indicated in Table 15.5. Averaging across the observed timeframe, the 436 AMXS had approximately 113 net effective technicians in its viable resource pool on any given day. This figure is approximately 27% of the total assigned quantity of technicians, again using the previously discussed definition for technicians.

Therefore, to arrive at the results shown in Table 15.5, the study team considered the factors from Tables 15.1 and 15.2, as well as the ancillary and CBT factors complimentary effectiveness rates calculated.

Each factor and rate detailed to this point was assigned a new designation for ease of use in the proposed NEP equation. The newly designated factors, factor descriptions, and the associated values are listed in Table 15.6.

The *T* factors relate to training, the *A* factors relate to available personnel, and the *P* factors relate to productivity. These factors were applied to the number of available technicians as recorded in the AMXS availability snapshots using the newly proposed NEP calculation, shown as Equation 15.1. Equation 15.1 is the cumulative NEP equation which accounts for all three factors which create variability in the resource pool and yields a numerical quantity of net effective personnel. To determine the NEP percentage, one need simply divide the right side of the equation by the number of assigned technicians (7-level nonmanagers, 5-levels, and 3-levels).

Figure 15.3 provides an Excel spreadsheet snapshot of an example NEP calculated for a generic maintenance unit. The maintenance unit's NEP is calculated using Equation 15.1 by entering the personnel totals in each of the five categories in the left column. These values are then multiplied by the factors in the right column to determine NEP. In this example, the unit has 104 technicians available but the NEP is only 77. In other words, the practical available maintenance capacity is only 77 technicians, not 104 as it initially appears.

To summarize, the study team's arrival at NEP followed an iterative sequence of three factor reductions:

- Skill-level productivity differences, to include those for trainees and trainers.
- Ancillary training and CBT.
- The nonavailability of personnel.

$$\mathrm{NEP} = T_{75}(A_{75NT} + (P_t A_{75T})) + T_3(P_e A_3) \tag{15.1}$$

TABLE 15.5 436 AMXS NEP Descriptive Statistics

	March 2007				April 2007				March-April 2007			
411 Assigned	Min	Max	Mean	Range	Min	Max	Mean	Range	Min	Max	Mean	Range
Available	79	167	120	88	77	124	105	47	77	167	113	90
% of Assigned	19%	41%	29%	21%	19%	30%	26%	11%	19%	41%	27%	22%

TABLE 15.6 NEP Factors

Factor	Description	Value
T_{75}	Ancillary/CBT factor for 7- and 5-levels.	0.948
A_{75NT}	The number of available nonmanager 7-levels and 5-levels who are not trainers.	Varies day-to-day
P_t	Trainer productivity.	0.85
A_{75T}	The number of available nonmanager 7-levels and 5-levels who are trainers.	Varies day-to-day
T_3	Ancillary/CBT factor for 3-levels.	0.925
P_e	Trainee productivity.	0.4
A_3	The number of available 3-levels.	Varies day-to-day

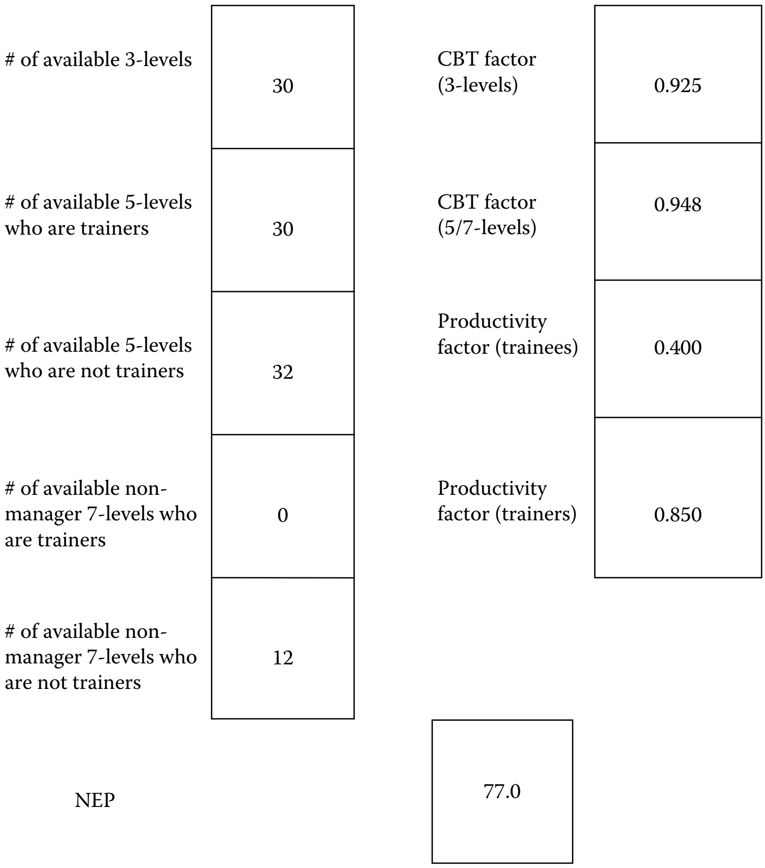


FIGURE 15.3 Example NEP calculation.

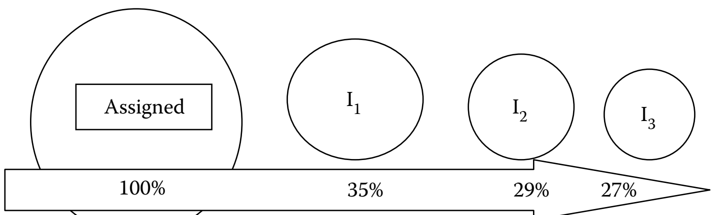
Figure 15.4 graphically illustrates these iterations based on the relative size of the impact of the three factors on reductions to the overall resource pool. As shown in Figure 15.4, nonavailability had the biggest impact, productivity factors were next, and finally the effect of CBT and ancillary training had the smallest impact.

In addition to AMXS, an Air Force Maintenance Group usually includes a separate equipment maintenance squadron (EMS) and component maintenance squadron (CMS). However, if total authorizations are under 700, EMS and CMS will be combined into a maintenance squadron such as the MXS at Dover AFB. Various flights within a typical MXS maintain aerospace ground equipment, munitions, off-equipment aircraft and support equipment components; perform on-equipment maintenance of aircraft and fabrication of parts; and provide repair and calibration of test, measurement, and diagnostic equipment [25]. Technicians assigned to MXS usually perform maintenance not explicitly linked to the launch and recovery of aircraft (as is the focus of AMXS).

However, some MXS personnel directly support flight line activities.

A more complete representation of the net effective personnel pool for aircraft maintenance resources in an MXG would include not only personnel in AMXS, but also those in MXS. The number of nonmanager 7-levels, 5-levels, and 3-levels assigned to the 436 MXS was determined from Air Force Personnel Center data to be 318 [26]. Using the study team's definition of technician, this results in 729 technicians in the 436 MXG (411 in AMXS plus 318 in MXS). However, because the study team could not obtain exact daily availability figures for MXS similar to those of AMXS, the study team applied each of the calculated daily NEP percentages for AMXS against the number of assigned technicians to MXS. This calculation yielded daily estimates of the number of NEP for MXS. Since AMXS and MXS are both aircraft maintenance units with many of the same AFSCs and similar demands on their personnel, any differences from actual numbers as a result of this method were considered negligible for this analysis.

The study team then added the AMXS NEP figures to the MXS NEP figures, resulting in a collective NEP figure for the flight line maintenance at Dover AFB. These collective NEP figures are shown in Table 15.7. The upper portion of the table shows the NEP figures grouped by columns (day of the week) with each row representing 1 or the 9 weeks over the entire period that data was tracked. The bottom section of Table 15.7 also displays the descriptive statistics for NEP across both AMXS and MXS combined. The highest average NEP value was 222 on Thursdays, representing approximately 30% of the baseline total of 729 people.



- Iteration 1 (I_1): Availability
    - $A_{75NT} + A_{75T} + A_3$
- Iteration 2 (I_2): Availability and productivity
    - $A_{75NT} + P_t A_{75T} + P_e A_3$
- Iteration 3 (I_3): Availability, productivity, CBT and ancillary training
    - $T_{75}\,(A_{75NT} + P_t A_{75T}) + T_3\,(P_e A_3)$

FIGURE 15.4 The iterations of NEP.

TABLE 15.7 Day of the Week NEP Distributions for 436 MXG (AMXS and MXS)

	Day of the Week NEP Distributions						
	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
NEP	186	219	228	211	259	219	187
	148	209	226	219	213	182	140
	153	212	211	242	219	195	155
	188	242	289	297	245	205	169
	165	210	220	216	294	235	198
	137	186	187	195	205	175	148
	173	206	192	188	194	176	168
	167	213	201	195	183	186	174
	176	203			185	194	180
n	9	9	8	8	9	9	9
Min	137	186	187	188	183	175140	140
Max	188	242	289	297	294	235	198
Mean	166	211	219	221	222	196	169
% of Assigned	23%	29%	30%	30%	30%	27%	23%
Range	51	56	102	109	110	59	58
Variance	300	221	1031	1241	1385	404	349
Standard deviation	17	15	32	35	37	20	19

Values in are rounded to nearest whole number.

15.9 Conclusion

The ratio between authorized and assigned personnel is typically used to quantify personnel availability. While this ratio is an indicator of maintenance capacity, it provides on a limited amount of information. These ratios do not take into account the abilities and skill levels of the maintenance personnel, nor does it factor in the availability of the personnel on a day-to-day basis. The NEP methodology described in this chapter is a repeatable process which produces NEP figures that provide leadership with a better representation of the personnel resources and actual capacity available to an Air Force aircraft maintenance organization on a day-to-day basis. The NEP methodology will be tested further and validated using personnel data from other units to verify similar results and potential gains. Ultimately, the NEP methodology has the potential to be used alone or in conjunction with LCOM to better portray maintenance personnel requirement and capabilities based on experience and skill levels.

As previously mentioned, the NEP methodology described in this chapter was developed as part of the larger C-5 TINMCM Study II. The entire study can be found at the Defense Technical Information Center Private Scientific and Technical Information Network website at https://dtic-stinet.dtic.mil/.

References

1. General T. M. Moseley. CSAF's Vector: Air Mobility's Strategic Impact, 6 June 2007. Available online: http://www.misawa.af.mil/news/story.asp?id=123055986, accessed 22 June 2007.
2. Honorable M. W. Wynne, and General T. M. Moseley. Strategic Initiatives, presentation to the House Armed Services Committee, 24 October 2007.
3. M. W. Wynne. Savings Less Than Expected in Drawdown, briefing to the Center for Strategic and Budgetary Assessments, Air Force Times. Available online: http://www.airforcetimes.com/news/2007/09/airforce wynnedrawdown 070924, accessed 9 September 2007.
4. AFI 21-101, Aircraft and Equipment Maintenance Management, HQ USAF/A4MM, 29 June 2006.
5. Study team notes from 439 MXG daily production meeting, Westover ARB, 12 December 2006.

6. Study team notes from meeting with 436 MXG/CD, Dover AFB, 8 January 2007.
7. Study team notes from meeting with 105 MXG/CC, Steward ANGB, 18 January 2007.
8. C. J. Dahlman, R. Kernchner, and D. E. Thaler. *Setting Requirement for Maintenance Manpower in the US Air Force.* Santa Monica, CA: RAND Corporation, 2002, 136.
9. J. R. T. Arnold, and S. Chapman. *Introduction to Materials Management*, 5th edn. Pearson Education, Lebanon, Indiana, 2003, 244.
10. J. G. Drew, K. F. Lynch, J. Masters, R. S. Tripp, and C. R. Roll, Jr. *Maintenance Options for Meeting Alternative Active Associate Unit Flying Requirements.* Santa Monica, CA: RAND Corporation, MG-611-AF, 2008.
11. Ibid.
12. Ibid.
13. Ibid.
14. Dahlman et al. 132.
15. This figure represents data from the 436 AMXS for March–April 2007.
16. Ancillary Training, Airman's Roll Call. Available online: http://www.af.mil/shared/media/document/AFD-070209-083, PDF, accessed 9 February 2007.
17. Ibid.
18. Study team notes from meeting with 436 MOS Supervision, 9 January 2007.
19. 436 AMXS/MXAA, Dover AFB, March–April 2007.
20. Education and Training Course Announcement, ETCA website. Available online: https://etca.randolph.af.mil/, accessed 21 March 2007.
21. Ibid.
22. US Office of Personnel Management, Federal Holidays. Available online: http://www.opm.gov/fedhol/, accessed 11 April 2007.
23. 436 AMXS/MXAA, Dover AFB.
24. Ibid.
25. Ibid.
26. Air Force Personnel Center, data pull, 27 March 2007.

IV

Contingency Planning and Logistics

16

Joint and Multinational Campaign Planning: A Project/Program Management Approach

Richard F. Deckro
Air Force Institute of Technology

James T. Moore
Air Force Institute of Technology

Michael L. Fredley
National Reconnaissance Office

Jack A. Jackson
Institute for Defense Analyses

Michael J. Artelli
Air Combat Command

John C. Van Hove
Air Education and Training Command

16.1 Background

Campaign planning is not only a complex and intriguing area of military study; it is potentially a matter of national survival. A nation may boast of its technological edge or of the training, determination and courage of its soldiers, but no degree of technology or heroism will protect a nation without an effective plan for their use. For military operations of the 21st century, coordination and cooperation between sister services, coalition and allied forces is a primary concern. It is this need for coordination and cooperation between services and partners, which underscores the need for joint and multinational planning. This chapter presents a project/program management framework for joint and multinational campaign planning that may prove useful for the effective use of the resources the United States has at its disposal.

The Project Management Institute defines a project as "A temporary endeavor undertaken to create a unique product or service" [11] while a program has been described as a group of projects executed over an extended period of time [1]. These are exactly the conditions that involved in planning, scheduling, resourcing and executing the planned engagements (projects) of a campaign (program). The techniques used in project management are easily applied to campaign planning for joint and coalition forces. Nicholas confirms this relation with his reference to the D-Day landings as perhaps the ultimate project [10].

To better understand joint and multinational campaign planning, it is helpful to understand where it lies in the strategy process. The President of the United States is responsible for developing the national strategic direction. Several documents are used to communicate the strategic direction to the Department of Defense. These include, but are not limited to, the National Security Strategy, National Strategy for Homeland Security, and the National Defense Strategy. These strategy documents are used to create the national strategic objectives. The Chairman of the Joint Chiefs of Staff (JCS) develops the National Military Strategy in response to the President's strategic direction, through the Joint Strategic

Planning System. The Combatant Commander (CCDR) then formulates the theater strategy through the Joint Operation Planning and Execution System (JOPES) [6]. At the campaign level, joint campaign planning is a part of joint operation planning and is notably affected by military strategy and ultimately national security objectives. As an illustration, note how closely the theater objectives of the Gulf War paralleled the national objectives of the war.

Gulf War national objectives:

1. Immediate, complete, and unconditional withdrawal of all Iraqi forces from Kuwait.
2. Restoration of Kuwait's legitimate government.
3. Security and stability of Saudi Arabia and the Persian Gulf.
4. Safety and protection of the lives of American citizens abroad.

Translated into theater/campaign objectives:

1. Attack Iraqi political/military leadership and command and control.
2. Gain and maintain air superiority.
3. Sever Iraqi supply lines.
4. Destroy Iraqi nuclear, biological, and chemical (NBC) capability.
5. Destroy Iraqi Republican Guard forces.
6. Liberate Kuwait City [5].

Reduced to its essentials, successful joint and multinational campaign planning requires that a CCDR considers the following:

1. The law of war, implementation of national policies, and protection of US citizens, forces, and interests.
2. Integration of deterrence measures and transition to combat operations.
3. Adjustments for multinational, interagency, or NGO circumstances.
4. Identification of termination criteria.
5. Identification of potential military requirements across the range of military operations.
6. Support for security assistance or nation assistance.
7. Inputs to higher strategies or subordinate planning requirements [8].

Leonhard [9] points out that today's warfare is both a spatial and a temporal affair requiring the coordination of a vast array of resources. Throughout the ages, the coordination of vast arrays of resources and personnel has been a military function. The ancient cry, "All roads lead to Rome" reverberated through the ages because the legions built the roads to serve that function. The basic principles of program and project management can trace their ancestry to campaign planning, while modern program management concepts descend in a direct line from World War II weapons programs and the Cold War's aerospace efforts [10].

This chapter looks at bringing these concepts of program and project planning full circle back to campaign planning. The project/program management framework presented in this chapter focuses on the sequencing of actions and on the application of resources in joint and multinational campaign planning and reflects the dependence of joint and multinational campaign planning on the higher levels of strategy. The framework also shows that joint and multinational campaign planning can impact higher levels of strategic planning as well.

16.2 The Joint and Multinational Campaign Planning Process

In order to motivate a need for an analytical framework for joint and multinational campaign planning, we must first consider two aspects of the planning process: the view of resources and the amount of time available for planning.

16.2.1 The View of Resources

Any discussion of campaign planning must consider resources. Resources, as used here, include weapon systems, personnel, supplies, spare parts, and any other material used in the course of military action. It is a fair assumption that joint and multinational campaign planning would be much simpler if a theater commander had unlimited, homogenous resources at his bidding. Unfortunately, resources are not unlimited and are, in fact, becoming scarcer with the reductions in each new defense budget. In addition, in a multinational setting, interoperability and compatibility of both men and material must be a consideration.

Depending on the commander's view and availability of resources, joint and multinational campaign planning can be conducted on a requirements planning basis or on a capabilities planning basis. From the theater commander's perspective, requirements planning can be summed up as: "This is the enemy threat and my assigned task. Here's my planned response. The resources I need for my planned response are X, Y, and Z". The problem with this approach is that resources X, Y, and Z may not be available at the levels the commander needs for his plan.

Capabilities planning, by contrast, can be summed up as: "This is the enemy threat and my assigned task. Here are the resources currently at my disposal. Given these resources, my planned response is W". This approach, too, has deficiencies. The resources at the commander's disposal may be totally insufficient to accomplish his/her assigned task no matter how skillful a planner he/she is or how valiantly his/her troops perform.

This idea was demonstrated in Desert Shield/Desert Storm. At the outset of hostilities by Iraq toward Kuwait, US Central Command (USCENTCOM) had insufficient in-house resources to conduct Desert Shield/Desert Storm. Consequentially, an execution based on capabilities alone would have resulted in total failure. Planning based on requirements, by contrast, seems more appropriate in this case (since resources can be drawn from other theaters), but even then, the planning cannot be conducted without considering the limitations of resources worldwide nor without weighing the consequences to other theaters. An analytical planning framework, therefore, must find a balance between requirements and capabilities and assure that the resources are available in the right quantities, at the right times, and at the right places.

16.2.2 The Amount of Time Available for Planning

Joint operation planning consists of two primary categories: contingency planning and crisis action planning (CAP) [8]. "While contingency planning normally is conducted in anticipation of future events, CAP is based on circumstances that exist at the time planning occurs" [7]. Figure 16.1 illustrates how campaign planning consists of contingency planning and CAP, and is within joint operation planning. Joint operation planning by the CCDR is a response to the joint strategic planning. Additionally, Figure 16.1 lists the products of contingency planning and CAP.

Planning for a Russian invasion of Europe is an example of contingency planning. "Contingency planning begins when a planning requirement is identified…, and continues until the requirement no longer exists" [7]. Planning Desert Shield/Desert Storm, on the other hand, is an example of CAP. Certainly, the Combatant Commander-in-Chief (CINC) of USCENTCOM, General Schwarzkopf, had conducted contingency planning for possible contingencies in the Persian Gulf region, but the exact nature and objectives of Desert Shield/Desert Storm were not known until Iraq actually invaded Kuwait (current doctrine would refer to General Schwarzkopf as CCDR, and more specifically COCOM since he had command authority). It was at that point alone that CAP could begin, perhaps using the plans developed during contingency planning as a starting point.

Both contingency planning and CAP use the Joint Operation Planning Process (JOPP) as a framework for the process. "JOPP underpins planning at all levels and for missions across the full range of military operations" [7]. The JOPP consists of seven steps as presented in Figure 16.2.

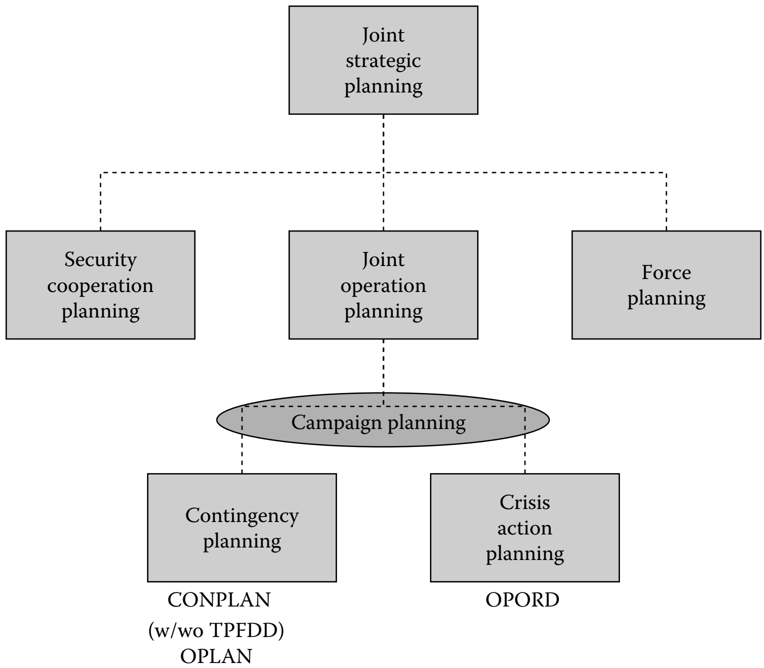


FIGURE 16.1 The campaign planning process.

The process for contingency planning assumes at the outset that the JCS has determined an allocation of resources to give to a specific theater commander, who in turn uses capabilities planning to develop his strategic concept and ultimately an OPLAN. The initial resource-allocation assumption is acceptable for the commander's purposes, but it says nothing of how the JCS arrived at that allocation of resources in the first place. At the very least, the initial apportionment of resources to any theater commander must also consider the needs of the other theaters of operation. The initial apportionment should also be related in some way to the actual requirements of the commander's campaign plans. Consequently, the apportionment of resources by the JCS to a theater commander should look laterally to other theaters and should also look in depth to the actual requirements of that theater. Here, an analytical framework could support the JCS planners.

CAP is perhaps more interesting to study than contingency planning since plans developed in a crisis are more likely to be executed. We cannot forget, however, that the plan developed in a crisis is greatly influenced by previous contingency planning since it is the contingency planning that has determined the current apportionment of resources around the globe and even the disposition of resources within the campaign theater.

In CAP, the CCDR shifts from a capabilities planning mode to a requirements planning mode. In the requirements planning mode, any new apportionment of resources by the JCS to the CCDR (beyond those resources already in the theater) is, by definition, responsive to the requirements of the CCDR's campaign plan. An additional apportionment of resources should, however, still be sensitive to the needs of other theaters. Hence, the same philosophy used in contingency planning (looking laterally and then in depth) should still apply in a crisis. As will become evident in the next section, the same analytical framework that could support contingency planning can respond to planning in a crisis.

16.3 A Project/Program Management Framework

The analytical framework presented here uses a project management approach, or more precisely, a *program* management approach [13–16]. The distinction made here between a project and a program

Initiation	Description/Products
Mission analysis	- President, SecDef, or CJCS decides to develop military options

Course of action (COA) development	- Revised staff estimates
	- COA alternatives including:
	    - Tentative task organization
	    - Deployment concept
	    - Sustainment concept

COA analysis and wargaming	- Potential decision points - Governing factors - Potential branches and sequels - Refined COAs - Revised staff estimates
COA comparison	- Evaluated COAs - Recommended COA - COA selection rationale - Revised staff estimates
COA approval	- COA modifications - JFC's COA selection - Commander's estimate - Refined commander's intent
Plan or order development	- Force planning - Support planning - Nuclear strike - Deployment planning - Shortfall identification - Feasibility analysis - Refinement - Documentation - Plan review and approval - Supporting plan development

FIGURE 16.2 Joint operation planning process. (From Department of Defense. 2006. *Joint Operation Planning*, JP 5-0. Joint Chiefs of Staff, Washington, DC.)

is simply a matter of magnitude, where a program might, in fact, be composed of a number of projects. The framework enables planners to decompose large problems, such as the apportionment of resources to theater commanders, into smaller, more easily managed *subproblems*. The subproblems may similarly be decomposed into even smaller subproblems. This allows the framework to adapt to virtually any specified degree of resolution. This same hierarchical planning framework is already used in the development and execution of weapons systems planning, development, and acquisition programs.

The ability to decompose large problems within this framework hinges on the structure of the problem. When appropriately formulated, many large problems, in general, and the joint and multinational campaign planning problem in specific, demonstrate a block-angular structure most frequently associated with the Dantzig–Wolfe decomposition algorithm [3]. This block-angular structure is depicted in Figure 16.3.

In Figure 16.3, we see a block-angular structure with three distinct blocks (A, B, and C), but the decomposition approach is not limited in the number of blocks it may have. The blocks represent sets of resource constraints that have only a subset of the problem's decision variables within them. Furthermore, any decision variable represented in a given block will be absent from all other blocks.

The block-angular structure may also have a set of coupling constraints, or constraints with decision variables that are represented in more than one block. To the right of the coupling constraints and the blocks are resource limits that are the right-hand side constants of the constraints.

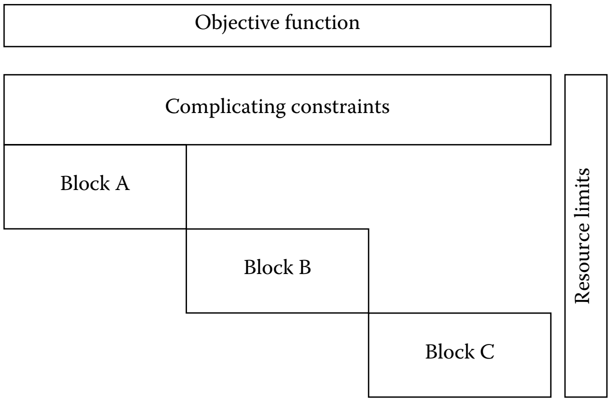


FIGURE 16.3 Block-angular structure.

Finally, the block-angular structure has an objective function. In a program management context, the objective function may be to minimize program costs or to minimize the program completion date.

As one expands any of the *blocks* within the hierarchy, the same block angular structure is repeated. In addition, within a specific block, or subproblem, an underlying network structure, given by the precedence relations, exists. Both of these elements provide opportunities to exploit the problem structure.

Utilizing the block angular structure of the planning problem, the problem may be decomposed. Wiley [14] and Wiley et al. [15] have shown how, at the aggregate level, the classic Dantzig–Wolfe approach can be used to effectively allocate resources among projects within a program. This same aggregate approach can be applied to allocate resources among CCDRs spread across several theaters. The work of Baumol and Tibor [2] suggests a methodology to trade-off the resources between commands to the benefit of the overall national objectives [4].

Within a theater, however, resources can not always be allocated on a continuous basis. An aircraft or tank simultaneously allocated to two points along the Forward Line of Own Troops (FLOT) will be unable to serve either tasking. Again, borrowing from the industrial planning sector, this is a classic aggregation/disaggregation found in the aggregate production-planning problem. The theater level problem of the CCDR has the same block angular structure, although its subproblem may represent battlegroups, armies, and wings. At this level in the hierarchy, however, integer decision variables must be considered. This suggests the application of Sweeney–Murphy Decomposition [12] to the integer block angular model shown in Figure 16.4.

The following mathematical program exists at *any* level of the hierarchy that one chooses to enter the process.

General model;

$$\text{Minimize} \quad z = c_1x_1 + c_2x_2 + \dots + c_px_p$$

Subject to

$$\begin{aligned}
A_1x_1 + A_2x_2 + \dots + A_px_p &= b_0 \text{ (Command)}\\
B_1x_1 \qquad\qquad\qquad &= b_1 \text{ (Sub-command 1)}\\
B_2x_2 \qquad\qquad &= b_2 \text{ (Sub-command 2)}\\
\dots\dots \qquad &\\
B_px_p &= b_p \text{ (Sub-command p)}\\
x_k \geq 0 \text{ and integer for } k = 1, \dots, p.&
\end{aligned}$$

where $B_k = \begin{bmatrix} F_k \\ G_k \end{bmatrix}$, $x_k = \begin{bmatrix} y_k \\ s_k \end{bmatrix}$, $b_k = \begin{bmatrix} f_k \\ g_k \end{bmatrix}$.

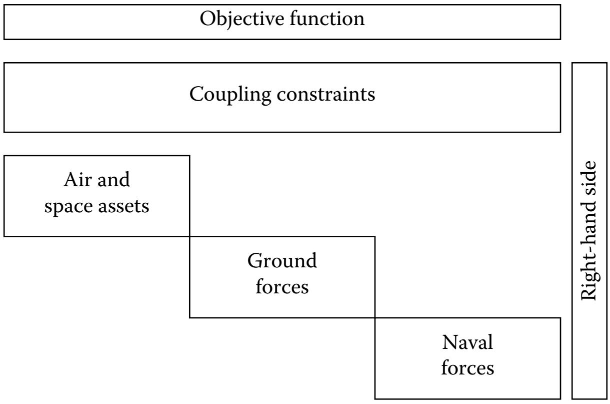


FIGURE 16.4 Theater level forces.

Matrix F_k and variables y_k are associated with resources while matrix G_k and variables s_k are associated with activity precedence, the sequencing of operational requirements. The right-hand side vector f_k and g_k represents *functions* of resource availabilities and durations, respectively. These may be linear, nonlinear, continuous, or integer. The resulting sub-problems, however, also produce an exploitable structure. The subproblem formation follows the traditional approach without additions:

$$\text{Minimize } z = (c_k - \pi \cdot a_k) \cdot x_k$$

$$\text{Subject to } F_k y_k \leq f_{k0} \tag{16.1}$$

$$-s_i + s_j \geq g_{k,(i,j)} \quad \text{for all (i, j) in project k} \tag{16.2}$$

$$LB_{k,(i,j)} \leq f_{k,(i,j)} \leq UB_{k,(i,j)} \quad \text{for all (i, j) in project k} \tag{16.3}$$

$$gLB_{k,(i,j)} \leq g_{k,(i,j)} \leq gUB_{k,(i,j)} \quad \text{for all (i, j) in project k} \tag{16.4}$$

$$x_k, y_k, s_k \geq 0$$

where c_k = the objective function coefficients associated with project k; a_k = the coupling constraint coefficient matrix associated with project k; x_k = the vector of variables associated with project k; and π = the shadow prices associated with the coupling constraints in the master problem.

Constraint type 16.1 represents any sub-command (or project) wide resources. Constraint type 16.2 is the precedence relations and maintains the underlying structure. Constraint type 16.3 and 16.4 maintain the bounds. Recall that the $f_{k,(i,j)}$ terms are functions which specify resource usage which in turn imply duration. They would have their own variable sets determining the amount of "accelerating" or "extending" that would occur. For example, for some activity (i, j) in sub-command k, the activities duration might be given as the multimodal function with R possible resource options [13]:

$$f_{k,(i,j)} = r_{k,(i,j),1}\, y_{k,(i,j),1} + r_{k,(i,j),2}\, y_{k,(i,j),2} + \ldots + r_{k,(i,j),R}\, y_{k,(i,j),R} \tag{16.R1}$$

$$y_{k,(i,j),1} + y_{k,(i,j),2} + \ldots + y_{k,(i,j),R} = 1 \tag{16.R2}$$

where $r_{k,(i,j),r}$ = the amount of resource used for option r allocation and $y_{k,(i,j),r} = 1$ if option r is selected and 0 otherwise.

For example, a movement may have different durations depending on the number and type of transport used. Constraint 16.R1 defines the duration according to which of the r resource levels are selected. Constraint 16.R2 is a special order set of type 1 that assures only one resource level is selected for the activity. These standard multimodal relations would apply to each appropriate duration.

At a given resource level, an activity duration will be dictated. This is given by the relation:

$$g_{k,(i,j)} = d_{k,(i,j),1}\, y_{k,(i,j),1} + d_{k,(i,j),2}\, y_{k,(i,j),2} + \ldots + d_{k,(i,j),R}\, y_{k,(i,j),R} \tag{16.D1}$$

where $d_{k,(i,j),r}$ = the duration of (i, j) option r allocation of resource. (As noted above, other types of relations for $f_{k,(i,j)}$ and $g_{k,(i,j)}$ may be considered.)

The decomposition adds another key advantage. The subproblems at each iteration must be solved. This offers an opportunity to consider parallel processing options. This would be effective in the contingency planning phases, and if the resources were available to the appropriate commanders, in the reactive phase of the process.

When a large linear program can be formulated in such a way that it demonstrates a block-angular structure, the Dantzig–Wolfe decomposition algorithm (Dantzig and Wolfe [3]) may be used to solve it more efficiently than if solved as a single linear program. The structure allows the model to be broken into smaller, manageable subproblems, a type of "divide and conquer" approach to seeking a solution. Additionally, the decomposition of the problem lends itself to certain interpretations not available in a linear program (see Baumol and Fabian [2]). As we describe the application of this decomposition to joint and multinational campaign planning, this interpretation will be made evident.

We have already seen that any analytical framework for joint and multinational campaign planning should have the following features:

(1) **Feature 1.** It should aid in developing a plan that will engender coordination and cooperation between services toward common objectives.
(2) **Feature 2.** It should allow the theater commander to appropriately sequence his chosen military actions and apply his resources to the accomplishment of those actions.
(3) **Feature 3.** It must find a balance between requirements planning and capabilities planning.
(4) **Feature 4.** It should enable the JCS to look laterally (at other theaters) and in depth (at a theater commander's actual campaign plans) when apportioning resources to the theaters of operation.

We start with the fourth feature, which will define the first level of resolution.

The first step in casting the joint and multinational campaign planning problem in a program management framework is to define the first level of resolution. This level of resolution will be the most general and far reaching level—that of apportioning resources to theaters. Recall that the apportionment of resources is the problem of the NCA/JCS and involves the theater CCDRs. Each CCDR is represented as a distinct block as in Figure 16.5. A block represents the constraints unique to a given CCDR. More specifically, these blocks represent the CCDR's campaign plans. (Again, while the figure shows only three theaters, the model will incorporate as many theaters as needed.)

There is also a set of constraints, labeled NCA/JCS, which are common to all the CCDRs or which are the concern solely of the NCA and JCS. For instance, these constraints may reflect political concerns outside the realm of any single CCDR or could be certain national technical assets.

The objective function may be a measure of national security or of cost. The nature of the objective function is an area meriting serious consideration.

The economic interpretations of how the Dantzig–Wolfe decomposition algorithm would seek to solve this problem not only satisfies feature 4, but also feature 3, of a desirable analytical framework as previously outlined. The algorithm starts with each CCDR planning his campaign using a requirements planning approach. In other words, the CCDR is allowed to plan as if in isolation and as if resources were limited only by the constraints specific to his theater.

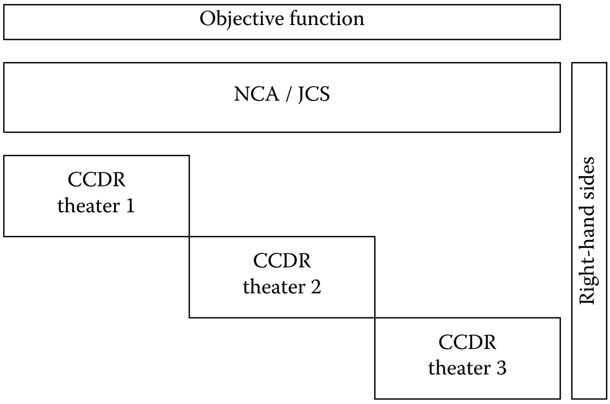


FIGURE 16.5 Apportionment of resources to CCDRs.

The NCA/JCS takes the CCDRs' campaign plans and seeks to modify them to best meet national objectives as specified in the objective function. The model can aid the NCA/JCS in deciding these modifications to the CCDRs' plans by identifying binding resources and, through duality relations, providing approximate measures of the relative benefit of the resources to each CCDR. Each CCDR, then, modifies his/her campaign plan to maximize the achievement of his theater objectives (the goal of his campaign plan) in terms of the revised benefits to the overall national objective [2,4].

This process continues iteratively until a set of plans is received from the CCDRs that maximizes the national objective. In this way, resources are apportioned in a way which maximizes the national objective (feature 4) and which creates a balance between the unconstrained realm of requirements planning and the often over-constrained realm of capabilities planning (feature 3).

A critical point that must be considered is that this process needs to be undertaken for many sets of world scenarios. While each CCDR may plan for his/her most difficult and resource-consuming campaign, each must consider contingencies should multiple Major Combat Operations (MCOs) occur concurrently. In the case of one MCO, we border on the current apportionment process. A framework like the one presented here will allow for a more balanced look at a one-MCO scenario since it can more explicitly consider the status quo requirements of the other theaters. It also allows for a more explicit look at other scenarios, whether they include two-MCOs, three-MCOs, or one-MCO with concurrent support to a humanitarian effort.

The second level of resolution is the CCDRs problem of campaign planning. This level differs from the first not only in its specific goals, but also in the way blocks are defined. If we draw a parallel to the first level, it might seem reasonable to let the blocks in this level represent service component commanders or functional commanders, all of whom work for the CCDR. This seems reasonable since the blocks in the first level are the CCDRs who work for the NCA/JCS and we think of the division of labor as stemming from "commander" to subordinate. At this level, however, it makes more sense for the blocks to represent operations of the campaign, since resource requirements are determined by the activities those resources support. We can go one step further. We can subdivide the operations into time frames.

Figure 16.6 depicts the timing of a hypothetical campaign's operations. Note that a time frame starts when an operation of the campaign either begins or ends. The lengths of the time frames are not necessarily identical. Time frame 1 may last for three days, for example, while time frame seven lasts for one month. The reason we subdivide operations into time frames is that the tasks that comprise an operation compete for resources with the tasks of concurrent operations. (We define tasks as specific activities conducted during an operation. If the operation is attacking command and control, then one task may be to jam a radar site and another may be to blow up a switching station.) In time frame 1, only three operations are competing for resources, whereas in time frame 3, six operations compete for the same resources. There should

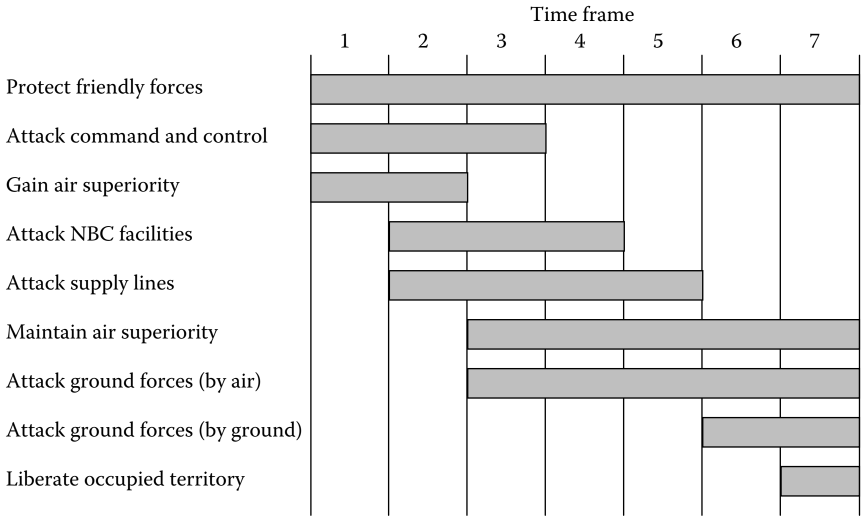


FIGURE 16.6 Timing of operations for the liberation phase of a hypothetical campaign.

obviously be a different apportionment of resources in time frames 1 and 3. In general, resources assigned to an operation in a time frame stay with that operation for the duration of the time frame.

One way resources have been allocated in the past is by assigning a percentage of the force to a specific function. For example, 40% of aircraft may be assigned to air superiority, 35% to interdiction, and 25% to close air support. This method would not work well in our hypothetical campaign since there is no close air support operations until time frame 6.

At the second level of resolution, then, we have a decomposition problem for each time frame. Figure 16.7 shows the block-angular structures for time frame 1.

An additional purpose of this level of resolution is to sequence, coordinate, and schedule operational tasks. Defining precedence relations between tasks or within and between operations does this. If we generically call the three operations in Figure 16.6 Operation A, Operation B, and Operation C, then our precedence relations may look like the network in Figure 16.8. With these precedence relations as part of the constraint set in Figure 16.6 and with an objective function to minimize the duration of the time frame, we will find an optimal apportionment of resources to each operation and the optimal sequencing and scheduling of tasks for each time frame.

As with the first level of resolution, there is an important economic interpretation of how the decomposition algorithm would seek to solve the CCDR's resource balancing problem. One way to view the process is to imagine that each operation has a "champion" whose main concern is the favorable outcome of his operation, regardless of the outcome of other operations. The champion for each operation is allowed to plan his operation as if in isolation, choosing what resources he wants and scheduling tasks in an optimal way.

The CCDR takes the operation plans from all the champions and seeks to modify them to best meet theater objectives as specified in the objective function. The CCDR seeks to modify the champions' plans by calculating a marginal benefit that each operation contributes to the overall theater objective. Each champion, then, modifies her/his operation plan to maximize the achievement of his objectives, but recalculated in terms of the revised marginal benefit to the overall theater objective.

The process continues iteratively until a set of plans, which maximizes the theater objective, is received from the commanders. Consequently, resources are apportioned and tasks scheduled in a way that maximizes the theater objective (feature 2). Since these imaginary champions bid for the resources which will best accomplish their goals without regard for the service component which owns them (like

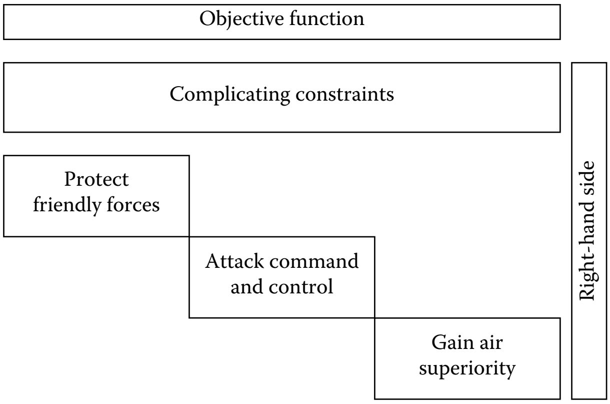


FIGURE 16.7 Apportionment of resources to operations during time frame 1.

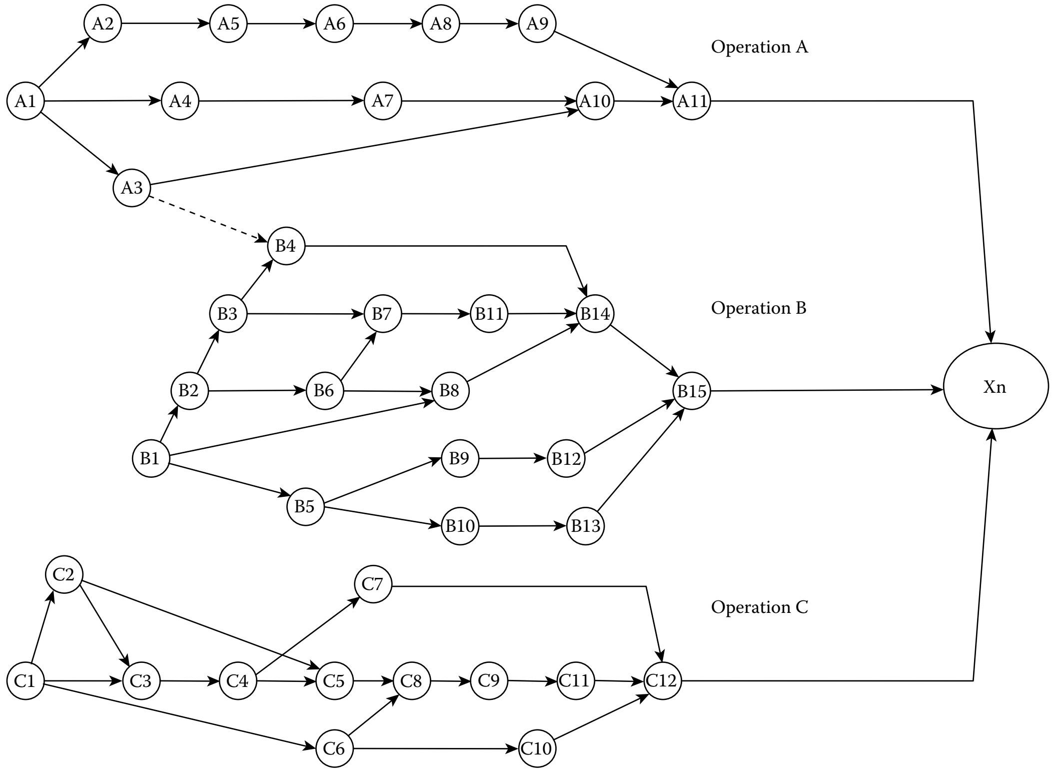


FIGURE 16.8 Precedence relations within and between three operations in time frame 1. (From Wiley, V.D. 1996. Optimization analysis for design and planning of multi-project programs. MS Thesis, AFIT/GOR/ENS/96M-18. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH.)

putting together a multiservice strike package), the process should promote greater coordination and cooperation between the services (feature 1).

When we finally mate both levels of resolution, we see the nested nature of this program management framework. The framework stresses the interdependence of the two levels. This implies not only that campaign plans depend on the apportionment of resources to theater commanders, but that the apportionment of resources to theater commanders should be influenced by the campaign plans.

In practice, the second level of resolution (the CCDR's problem) is transparent to the NCA/JCS. All the NCA/JCS knows is that the CCDR develops his campaign plans based on the incremental (marginal) benefits of the resources to the overall national objectives. Similarly, the first level of resolution (the NCA/JCS problem) is transparent to the CCDR's supporting commanders, who are primarily responsible for details of the campaign plan. All they know is that they must modify the campaign plan based on an incremental value of the resources passed to them by the CCDR. In a sense, the iterative process occurs between the NCA/JCS and the supporting commanders, with the CCDR as a negotiator.

16.3.1 Strengths of the Framework

(1) The framework is designed to complement, not replace, operational art in joint and multinational campaign planning. The tasks of defining objectives, assigning weapons to targets, determining courses of action and so forth are left to the military commander, not turned over to some machine. The framework acts as an analytical tool to help commanders manage resources more efficiently. The efficient management of resources ultimately influences the schedule and can make the commander more effective.
(2) The framework supports the NCA/JCS with more analytical information on apportioning resources to theaters of operation.
(3) A balance can be achieved between requirements planning and capabilities planning.
(4) The Dantzig–Wolfe decomposition algorithm provides an insightful interpretation to the consequences of allocating resources between theaters.
(5) Analytical methods derived from the decomposition approach can be applied to the joint and multinational campaign planning process.

Of course, as with all planning, nontangible factors, the unexpected, or just the "fog of war" will cause the actual execution to vary from the developed plan. This is as true in traditional project and program management as it is in campaign planning. The planning process, however, can be aided (and speeded) by the use of modeling and analysis. The planning process provides a baseline to analyze contingencies and requirements. It is also required to assure that proper capabilities and resources are developed and procured.

16.4 Summary

Utilizing a project/program management approach to aid in the establishment and coordination of joint and multinational operational planning can be a useful framework to support campaign planning. While the development of any campaign plan requires time, there are some immediate benefits from implementing the campaign as program analogy. A great deal of software currently exists for program and project management. This software is immediately available to study interdependencies, precedence relations, and timing in campaign planning. In addition, all of the research available for program and project resource planning can be brought to bear on the joint and multinational campaign planning problem. While the time interval may have changed from months and weeks in program planning to minutes and seconds in modern joint and multinational campaign planning, the planning and scheduling principles are the same. Whether in the days of Hannibal or Osama bin Laden, the tools of project management can be brought to bear on the problems of campaign planning and, of course, any multinational or multiorganization operation.

References

1. Angus, R.B., N.R. Gundersen, and T.P. Cullinare. 2002. *Planning, Performing and Controlling Projects.* 3rd Edition, Prentice Hall, Upper Saddle River, NJ.

2. Baumol, W.J., and F. Tibor. 1964. Decomposition, pricing for decentralized and external economies. *Management Science*, 11: 1–32.
3. Dantzig, G.B., and P. Wolfe. 1960. Decomposition principle for linear programs. *Operations Research*, 8: 101–111.
4. Deckro, R.F., V.D. Wiley, J.A. Jackson, M.L. Fredley, and J.C. Van Hove. 1998. Resource planning and coordination in multi-project programs. Working Paper, AFIT Center for Modeling and Simulation Analysis.
5. Department of the Air Force. 1995. *Squadron Officer School Course Manual (95-D).* Air University, Maxwell AFB, AL.
6. Department of Defense. 2007. *Doctrine for the Armed Forces of the Unites States*, JP-1. Joint Chiefs of Staff, Washington, DC.
7. Department of Defense. 2006. *Joint Operation Planning*, JP 5-0. Joint Chiefs of Staff, Washington, DC.
8. Department of Defense. 2006. *Joint Operations*, JP 3-0. Joint Chiefs of Staff, Washington, DC.
9. Leonhard, R.R. 1994. *Fighting by Minutes: Time and the Art of War*. Praeger Publishers, Westport, CT.
10. Nicholas, J.M. 1990. *Managing Business and Engineering Projects*. Prentice Hall, Upper Saddle River, NJ.
11. Project Management Institute. 2004. *A Guide to the Project Management Body of Knowledge* (*PMBOK® Guide*). Newtown Square, PA.
12. Sweeney, D.J., and R.A. Murphy. 1979. A method of decomposition for integer programs. *Operations Research,* 27: 1128–1141.
13. Van Hove, J.C. 1998. An integer program decomposition approach to combat planning. Ph.D. Dissertation, School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH.
14. Wiley, V.D. 1996. Optimization analysis for design and planning of multi-project programs. MS Thesis, AFIT/GOR/ENS/96M-18. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH.
15. Wiley, V.D., R.F. Deckro, and J.A. Jackson. 1998. Optimization analysis for design and planning of multi-project programs. *European Journal of Operational Research*, 107(2): 492–506.
16. Winnefeld, J.A., and D.J. Johnson. 1991. Command and control of joint air operations: Some lessons learned from four case studies of an enduring issue. RAND Report R-4045-RC. Santa Monica, CA.

17

Mobilizing Marine Corps Officers*

Dan O. Bausch
Naval Postgraduate School

Gerald G. Brown
Naval Postgraduate School

Danny R. Hundley
Naval Postgraduate School

Stephen H. Rapp
Naval Postgraduate School

Richard E. Rosenthal†
Naval Postgraduate School

17.1 Introduction

Almost all of the United States' contingency plans for responding with force to international crises involve rapid deployment of the Marines in the earliest phases of action. The Marines may be called upon to seize and hold a strategic geographic location or to negate a specific enemy asset. The exact mission will depend on the nature of the crisis, but in any case, it is essential for national security that the Marine Corps be able to rapidly mobilize its personnel from peacetime to wartime duties. We designed and built a system to assign marine officers appropriate duty assignments—or billets—during a crisis mobilization.

The officer assignment branch at Marine Corps headquarters is responsible for providing officers to billets if a mobilization occurs. The branch spends most of its time assigning officers' peacetime billets, but it occasionally engages in mobilization assignment exercises. In these exercises, a hypothetical crisis scenario is assumed, and the branch is supposed to go as far as printing (but not sending) MAILGRAM orders to report for officers to fill the required mobilization billets. The branch studies the time it takes

* This chapter was originally published in *Interfaces* journal: Reprinted with permission. Bausch, D.O., G.G. Brown, D.R. Hundley, S.H. Rapp, and R.E. Rosenthal. 1991. Mobilizing marine corps officers. *Interfaces*, 21(4), 26–38. This paper won the 1989 Koopman Award from the Military Operations Research Society.

† Posthumous.

to finish the exercise and evaluates the quality of the resulting officer assignments. The branch concluded from past performance that improvements were needed.

17.2 Problem Objectives

Since the officer assignment branch spends most of its time on peacetime billets and we are concerned here with mobilization billets, we will explain the differences between the two.

First of all, problem size and urgency differ greatly. In peacetime, active-duty marine officers receive new assignments about once every three years, whereas, during mobilization, all active-duty, reserve, and retired officers are eligible for immediate reassignment. In the words of the branch chief, mobilization requires "years" worth of work in a matter of days.

Secondly, the peacetime and mobilization assignment problems have different measures of effectiveness. In peacetime, the officer's career development and professional desires are major considerations. Each officer should amass a collection of skills and experiences that enhance the Marine Corps' long-term effectiveness. During mobilization, the Marines' purpose is much more straightforward: just fill the required billets with the best possible officers. In the urgency of mobilization, we can ignore officer development considerations. But we must carefully examine the skills officers currently possess and determine how and where they can best be deployed in the present crisis.

We address the officer mobilization problem with an optimization model that combines three objectives:

1. Maximize fill, that is, maximize the number of billets filled by officers with acceptable (or better) qualifications.
2. Maximize fit, that is, attempt to fill billets with officers whose qualifications are not merely acceptable but come as close as possible to fitting the billets perfectly.
3. Minimize turbulence, that is, try to keep officers assigned to the same unit they were assigned to before mobilization, or, failing that, try to have them reassigned to a nearby unit.

Our ability to model and measure these criteria varies. The *fill* criterion is defined simply as the percentage of billets filled, so it is easily measured. The *fit* criterion is subjective and requires an approximate model based on several criteria for matching officers to billets, including grade, sex, special training, and status (active, reserve, or retired). Turbulence is a lower priority criterion than fit or fill but is still very important. We define *turbulence* as the percentage of assigned officers whose mobilization billets requires them to report to a unit more than 100 miles away from their current assignment.

17.3 Previous Mobilization Methods

Prior to our work, the only tool the Marines had to help with mobilization assignment was the officer staffing goal model (OSGM) (Decision Systems Associates 1983). OSGM was designed to provide peacetime staffing targets. When it was created, it was not intended to provide mobilization assignments. Even so, the Marines relied on it in mobilization exercises for many years.

The Marines had several reasons for wanting a better mobilization system than OSGM:

1. Solution quality: OSGM focuses on peacetime factors that are irrelevant for mobilization and ignores things that are important, such as turbulence. Optimization (with a focus on mobilization issues) should produce better solutions.
2. Timeliness: At the time our project was undertaken, it took 2–4 days to complete a mobilization assignment exercise with OSGM. This was largely due to the fact that OSGM has to be run on a remote, leased computer. Undoubtedly, the Marines would like to be able to try several model runs before committing to action, so fast turn-around is important.

3. Cost: The Marines spend a great deal for external execution and maintenance of OSGM. Because the OSGM is written in machine-specific code for a 1970s vintage Cyber computer, it is very expensive to maintain and not transportable to more modern computers. Our in-house model residing on a personal computer is much cheaper to execute and maintain, and it has already been transported effortlessly between computing platforms.
4. Reliability: A mobilization system must work on the first try.

The Marines asked the Naval Postgraduate School to develop an improved system, first as a masters' thesis (Rapp 1987) and then as a faculty research project (Bausch and Brown 1988). We decided to take advantage of the 386-based personal computers that we had recently demonstrated to be capable of large-scale optimization and to exploit the suite of optimization software that was installed in the 80386 environment for this purpose (Bausch and Brown 1988).

The military has made use of optimization modeling for manpower planning in other instances (Gass et al. 1988; Grinold and Marshall 1977; Klingman et al. 1984; Klingman and Phillips 1984; Liang and Buclatin 1988; Liang and Thompson 1987). As far as we know, we are the first to specifically address officer assignment during mobilization.

17.4 Data and Terminology

Two files are crucial for our work. The wartime officer slate file (WOSF) contains detailed information on every officer. The wartime authorized strength report (WASR) describes every wartime billet for a mobilization scenario. Several versions of WASR are maintained for various war plans. The quick-response mobilization system crucially depends upon the Marine Corps's commitment to sustained, in-house maintenance of the WOSF and WASR data bases (Tables 17.1 and 17.2). We explain special terms as:

A *monitor command code* (MCC) designates the unit of a particular officer billet.

A *military occupational specialty* (MOS) is a four-digit code representing an area of expertise that requires specialized qualification and training. Some officers have earned a primary MOS (PMOS) plus one or two additional MOS's (AMOS).

A few of the MOS's in WOSF are catch-all codes for officers whose specialties are outdated. Similarly, some of the billets do not require special expertise and are coded with an imprecise MOS. We refer to these unspecialized billets as generalized billets and the others as regular billets. Some generalized billets are partially specialized in that they are restricted to ground officers or aviators.

The *staffing priority level* (SPL) of a wartime billet indicates its priority. The higher the SPL, the more crucial it is to fill the billet with an officer of the right fit.

The grades included in WOSF and WASR are warrant officers through colonels. Generals are omitted because their billets are preassigned.

17.5 Conceptual Network Model

Figure 17.1 shows a network model in which each officer in WOSF is represented by a node on the left-hand side and each billet in WASR is represented by a node on the right-hand side. In this conceptual network, the officer nodes have a supply of one and the billet nodes have a demand equal to the number of officers required.

If an officer is eligible for a billet, a directed arc connects the corresponding officer and billet nodes. Eligibility depends on the input data (Tables 17.1 and 17.2) and on numerous Marine Corps rules and policies (for example, no retired officers wanted in combat billets, no grade substitutions wanted in high-priority billets). The cost of an arc is a weighted sum of a measure of the quality of the officer-billets fit and the distance between the officer's current MCC and the billet's MCC (see Appendix in this chapter).

TABLE 17.1 Officer Supply Data

For Each Officer

(a) Social security number

(b) Grade

(c) Current monitor command code (MCC)

(d) Primary military occupational specialty (PMOS)

(e) First additional MOS (AMOS1)

(f) Second additional MOS (AMOS2)

(g) Officer type: regular, reserve, or retired

(h) Sex

(i) LDO (limited duty officer) status

The wartime officer slate file (WOSF) is a data base that contains current records on all active, reserve, and retired marine officers. Our mobilization system uses WOSF as input and extracts the listed attributes for all officers who are eligible for mobilization. Officers with matching attributes are temporarily aggregated into "officer supply nodes" for a network optimization model. The WOSF contains as many as 40,000 eligible officers, from whom aggregation yields about 10,000–15,000 supply nodes.

TABLE 17.2 Billet Demand Data

For Each Billet:

(a) Staffing priority level (SPL)

(b) Monitor command code (MCC)

(c) Grade

(d) Required MOS

(e) Number of officers needed

(f) Female officer allowed (yes or no)

(g) Limited duty officer allowed (yes or no)

The wartime authorized strength report (WASR) is a Marine Corps file that contains every required wartime billet for a specific mobilization scenario. The Marines maintain several versions of WASR for different war plans. Our system reads the listed billet attributes and temporarily aggregates matching billets into "billet demand nodes". A WASR file can contain as many as 25,000 billets, which are typically reduced about three-fold by aggregation.

There is a high probability that some billets will remain unfilled in any given mobilization because of a shortage of eligible officers. To account for this eventuality, the conceptual network has an extra node, called *clonemaker*, that represents a fictitious large supply of officers who can fill any billet at a very high cost. The conceptual model has an arc connecting the clonemaker node to all billet nodes.

There is also a very good chance that some officers (particularly retired officers) will not be eligible for any unfilled billets and, hence, will remain unassigned. To account for this possibility, an extra billet node called *unused* is added to the conceptual model, with explicit arcs connecting all officers' nodes to this node. The clonemaker and unused additions to the conceptual model guarantee network feasibility.

One of us (Rapp) implemented a prototypic version of the conceptual model using the NETSOLVE package (Jarvis and Shier 1988). This prototype gave encouraging results, but NETSOLVE could handle only a very small number of officers and billets compared to the needs of a real mobilization problem.

Our next implementation of the conceptual model (Rapp 1987) used the GNET network optimizer (Bradley et al. 1977). This implementation, dubbed MCMAM, yielded concrete improvement in solution quality over OSGM, for example, about 6% greater fill. MCMAM did not stand alone, it relied on the Statistical Analysis System (SAS Institute 1985) for reading, sorting, and error-checking the WOSF and WASR data bases. On an IBM 3033-AP mainframe, it took 5 minutes of SAS time and 30 minutes of MCMAM time to generate and solve a 27,000-officer, 10,000-billet problem. We deemed this computational performance inadequate to warrant converting the system to a personal computer or installing it at Marine Corps headquarters. Accordingly, we engaged in further research to improve performance.

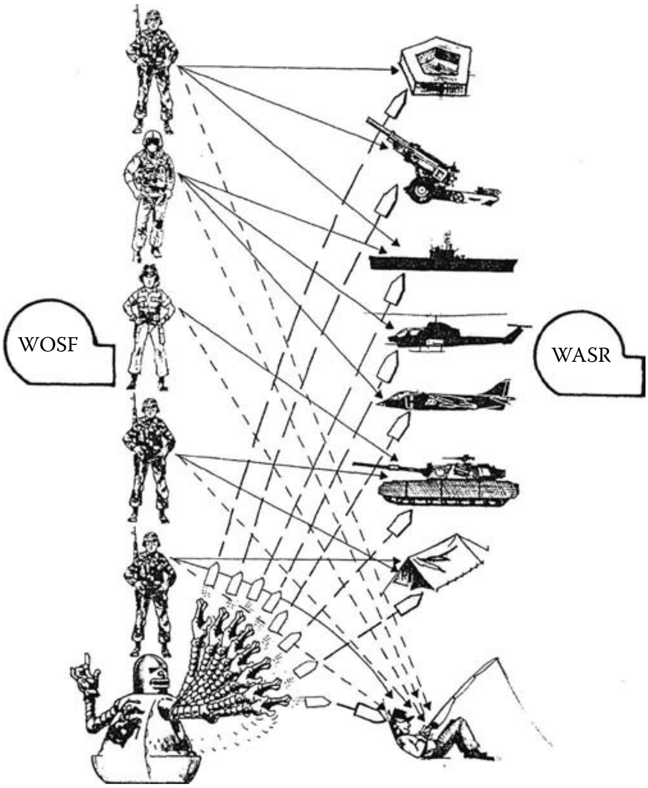


FIGURE 17.1 A conceptual network model of the Marine Corps mobilization problem depicts each officer as a supply node and each billet as a demand node. The "clonemaker" node at the lower left accounts for the possibility that some billets will remain unfilled due to a shortage of eligible officers. Conversely, the "unused" node at the lower right accounts for available officers who are not eligible for any unfilled billets. A literal implementation of the conceptual model would be computationally impractical, so our mobilization system employs several important refinements.

17.6 Practical Refinements to the Conceptual Model

The conceptual model has some inherent computational impracticalities, so the model we built for the Marines differs from it in a number of important ways. The differences have to do with making the network smaller, reducing the work required to generate it and reducing the time required to solve it. The key changes to the conceptual model are summarized below:

1. Node aggregation: The number of nodes is substantially reduced by a temporary node aggregation. Officers who match one another with respect to grade, sex, limited-duty status, type, occupational specialties, and MCC are merged into a single *officer-supply node*. Similarly, billets with matching data attributes are merged into *billet-demand nodes*. These aggregations can yield threefold reductions in the number of nodes yet sacrifice nothing in terms of solution quality.
2. Arc screening: A realistic scenario exhibits as many as 40,000 available officers and 25,000 required billets. A literal implementation of the conceptual model would require eligibility tests for 1,000,000,000 officer-billet pairs. Fortunately, in practice most pairs are ineligible, so we do not have to worry about solving billion-arc networks, but it is vital to be able to pick out the eligible pairs as efficiently as possible. We expended a great deal of effort in data structure design and programming for the arc generation routine to ensure that most of the ineligible officer-billet pairs are not considered explicitly.

3. Priority separation: The problem can be optionally separated into subproblems, based on billet priority. The first subproblem assigns only the highest priority billets, subject to very tight officer-billet fit restrictions. Subsequent subproblems successively admit lower priority billets and less stringent fit criteria. This approach reflects the preferences of the Marine Corps and does not detract from our results. (Originally, this option was mandatory. We allowed it to be bypassed in a later modification.)
4. Generalized billet heuristic: Because so many officers are eligible for generalized billets, they are very easy to fill. Yet, for the same reason, they necessitate the generation of a burdensome number of arcs in the conceptual network. We chose, therefore, to treat the generalized billets differently from the regular billets, using a simple greedy heuristic rather than the network optimization model.
5. ENET solver: By using an elastic network program, ENET, we omitted the explicit arcs representing unused officers and "clonemaker-filled" billets and handled them implicitly. This results in a substantial reduction in the number of arcs. This is possible because the ENET algorithm treats networks as inequality-constrained linear programs, in which a dynamic subset of the flow conservation constraints are binding at any given iteration. ENET also employs automatic basis aggregation, as described for the XNET variant of GNET in Bradley et al. (1977).

   The preceding refinements, individually and collectively, result in the generation of much smaller networks than the conceptual model. By using judiciously chosen data structures, we generate these networks extremely rapidly. The next refinement is an algorithmic device, which might be referred to as a type of linear programming pricing strategy, and which greatly reduces network optimization times.
6. Successive restrictions: When solving one of our network subproblems, we initially consider all the arcs representing perfect officer-to-billet fits eligible and all other explicit arcs temporarily ineligible. ENET optimizes first over this restricted set. Although the resulting solution is suboptimal in the network at hand, it is found extremely rapidly and furnishes ENET with a good starting point for solving another less restricted version of the original subproblem. In the second restriction, ENET optimizes over all arcs with penalty costs up to one-third the maximum arc penalty cost. ENET then starts from the solution to the second restriction and performs a final optimization in which all arcs are eligible. As you would expect, the perfect arcs are preferred, and large numbers of increasingly imperfect arcs have diminishing influence on the decreasingly restricted solutions. This modest refinement renders speed improvements of between three-and 20-fold.

The computational benefit of all these refinements is shown in Table 17.3.

TABLE 17.3 Model Refinement Computations

Version Date	Refinements Added	Network Generation and Optimization (Mainframe CPU Time)
9/87	Node aggregation	30 minutes
	Priority separation	
	Arc screening	
11/87	Generalized billet heuristic	3.5 minutes
	ENET solver	
4/88	Specialized data structures	9 seconds
	Successive restrictions	

Our refinements to the conceptual model were added in stages in research versions of the mobilization assignment system. This Table shows the effects of the refinements on computation time for one subproblem containing 27,003 officers and 10,441 billets. The research versions of the system were implemented on an IBM 3033-AP mainframe, whereas the version currently used by the Marines resides on a personal computer.

17.7 Implementation

Application of the preceding ideas leads to an efficient mobilization system. We developed research versions of the system on an IBM 3033-AP mainframe computer under CMS in VS FORTRAN (Table 17.3). We then implemented the system in NDP FORTRAN-386 (MicroWay 1988). (See Bausch and Brown (1988) for a complete description of this PC programming environment.) About 18 months later, we switched to the SVS FORTRAN 77 compiler (Silicon Valley Software 1990), which sped up the program by a factor to two. The Marines originally ran the mobilization system on a Compaq desktop personal computer with a 25-megahertz 80386 processor, 80387 coprocessor, and nine megabytes of memory. It now runs on 80486 PC as well. A run of the system proceeds as follows.

17.7.1 Step 1: Data Input and Node Aggregation

We read three input files: WOSF, WASR and a small file containing policy parameters that define the cost function and the eligibility rules. The WOSF and WASR files are read once and carefully checked for errors. Good records are aggregated and stored in a binary file. Bad records are excluded from the model and reported in exception files.

Step 1 takes almost half of the total time of a complete run of the system, but if there are multiple runs (for example, with different values for the policy parameters), it needs to be performed only once. The binary file contains pointers that are used later for disaggregation.

17.7.2 Step 2: Network Generation and Solution for High-priority Regular Billets

We generate an elastic network model that is restricted to the high priority regular billets and to the officers who can fill them. Then we call ENET as a subroutine and obtain an optimal solution. The optimal assignments are stored on another binary file, while officer availabilities and billet demands are updated accordingly.

17.7.3 Step 3: High-priority Generalized Billet Assignment

Each high-priority generalized billet is assigned to the closest available officer of the right grade, subject to sex, limited-duty and air/ground restrictions. These assignments are added to the binary output file and appropriate updates are made.

17.7.4 Step 4: Medium-priority Subproblem Generation and Solution

We repeat Steps 2 and 3, for regular and generalized billets, except now we restrict attention to medium-priority billets and any high-priority billets that remain unfilled.

17.7.5 Step 5: Low-priority Subproblem Generation and Solution

We repeat Step 2 and 3 for regular and generalized billets, except now we consider low-priority billets and any higher-priority billets that remain unfilled. After ENET solves the last subproblem, we produce a summary report on cumulative solution quality (similar to Table 17.4).

17.7.6 Step 6: Node Disaggregation and Solution Reporting

If the user desires, we create detailed reports on filled and unfilled billets. The optimal assignments are disaggregated to an individual officer-to-billet level, and are placed in a file which can be used as input to a MAILGRAM printing program.

17.8 Early Results

The outputs from many versions of our system have been carefully scrutinized with the view of revealing data deficiencies, modeling oversights, and programming errors. Preliminary criticisms have enabled us to identify previously unelucidated institutional policies (a frequent unadvertised benefit of applied operations research).

The approved solutions exhibit the qualities summarized in Table 17.4. Total computing time on the Marines' 80386-based personal computer is under 10 minutes, with the time divided among tasks as reported in Table 17.5.

The model run reported in Tables 17.4 and 17.5 uses a full-scale marine mobilization scenario. That problem could not be run on the old system used for mobilization, OSGM, because of its large size, but we have compared results on smaller problems. In every case, the new system achieves better quality solutions with respect to every measure of effectiveness considered.

TABLE 17.4 Mobilization Prioritization

	Priority			
	High	Medium	Low	Total
Number of billets	13,625	12,186	938	26,749
Percentage of billets filled	94.9	91.1	94.0	93.2
Percentage of filled billets in which assignment uses:				
Perfect grade fit	84.4	79.6	91.3	82.4
Perfect MOS fit	92.8	87.6	72.0	89.7
No turbulence	58.3	42.0	14.5	49.3
Active-duty officers	65.9	50.9	19.3	57.4
Reserve officers	19.6	25.1	9.9	21.8
Retired officers	9.4	15.1	64.9	14.0

The Marines are concerned about several measures of effectiveness in officer mobilization. The primary objective is to maximize the number of billets filled with suitably qualified officers. The second objective is to maximize the quality of officer-to-billet fit. Fit is evaluated with respect to several criteria, including grade fit, MOS fit, and preference for active-duty officers and reserves over retired officers. The third objective is to minimize turbulence, defined as the percentage of assigned officers whose mobilization billet required them to report to a unit more than 100 miles away from their current assignment. Results of our mobilization system for a full-scale marine mobilization scenario are reported. This example is too large to run on the Marines' old system; but, on smaller problems where comparisons could be made, the new system always produced significantly better results with respect to all measures of effectiveness.

TABLE 17.5 Computing Effort as Percentage of Total Time

Data input and node aggregation	48%
Network generation	33%
Network optimization	
Generalized billet assignment	
Node disaggregation and report writing	19%
	100%

Our mobilization system provides the Marines with sufficiently rapid response to be used in wartime. On a personal computer, it takes under 10 minutes for full-scale Marine Corps mobilization, with computational effort distributed as shown. The system can be run in two ways. In one option, separate networks are generated and solved for each priority level. Alternatively, the system can solve a single network encompassing all billets. Using the first option, the largest subproblem to date has about 21,000 nodes and 120,000 arcs. The largest problem encountered to date using the second option had the same number of nodes and over 1 million arcs.

17.9 Subsequent Results

In the first 18 months that our system resided at Marine Corps headquarters, it was used extensively, and it underwent some significant changes. Among other things, the system was named and renamed. First it was called OMAM, for officer mobilization assignment model, and then it became MARS, for manpower assignment recommendation system.

As often happens with the installation of an optimization-based system, the most significant outcome in the early applications was the discovery of errors in the input data. (Optimizers tend to hone right in on bad data, unlike simulations and statistical analyses, which tend to wash out their effects.) Dan Bausch was assigned to fix the errors while on temporary active duty as a reserve marine officer. He redesigned the input files so they are now much easier to understand, verify, and modify.

Bausch also added new information to the input files that enables MARS's network generator to comprehend and obey more complex eligibility rules than before. This results in better fit. For example, the billet file now specifies grade and MOS substitution policies for each billet individually, yielding new flexibility. Also, there are now matching "compatibility fields" in the officer and billet files, so, for instance, if a billet requires an officer with top-secret clearance, MARS enforces this restriction. Not all of the compatibility fields are currently used, so there is room to accommodate future considerations. In general, the input file structure and the eligibility logic are now sufficiently flexible to allow MARS to be used for peacetime as well as mobilization assignment. MARS has been tested on a limited basis in peacetime scenarios.

Bausch also added a great deal of reporting capability, which is another common occurrence in the early period of adoption of an optimization-based system. As the users learn more from and about the system, they tend to request new ways to summarize and present the results.

Two criticisms of the system emerged in the early going. One has been permanently rectified; the other may now be circumvented at the user's discretion.

The system originally aggregated monitor command codes into geographic regions, which made the node aggregations more effective. However, because of this geographic aggregation, our early system was criticized for inaccurately measuring the travel distance between an officer's current location and his mobilization billet, (particularly if he was moving within the same region). As a result, we dispensed with the geographic aggregation and use more accurate MCC-to-MCC distances in the evaluation of all potential assignments. Solution quality has improved as a result of this change but at the cost of a small increase in computing time. The officer node aggregation is now such that if two officers belong to the same node, the only difference between them in the WOSF date base is their social security number.

The second area of criticism involves priority separation. Strictly speaking, the critics are right in saying that this procedure potentially sacrifices some optimality. (Some Marine Corps manpower planners have a surprisingly devout attitude toward optimization.) In our view, the objective function of the optimization model is not meaningful per se, but rather is a compilation of many policies and preferences, among the most important of which is to fill the top priority billets first. In other words, priority separation is not only a computational convenience but also an accurate reflection of Marine Corps official policy. Though we still believe in this justification, we heard the criticism enough times to do something about it.

MARS now offers the option of omitting priority separation, thus optimizing all billet assignments in one very large problem. A recent instance of such a problem, involving 16,739 officers and 16,411 billets (a peacetime scenario), had the following performance characteristics:

Number of nodes = 20,942;
Number of arcs = 1,059,607;
Network generation time = 1.06 minutes; and
Network optimization time = 8.71 minutes

on a Compaq 486/33. In contrast, the same problem with priority separation takes one-sixth the time and requires much less computer memory. The solution obtained without priority separation has greater total fill but it sacrifices quality of fit in the high priority billets.

Whether or not priority separation is appropriate, we expect the option of circumventing it to be exercised frequently. It was quite comforting, therefore, to discover that our system can generate and solve problems with over a million variables in under 10 minutes on a personal computer.

17.10 Conclusions

United States' defense plans rely upon our ability to mobilize the Marines Corps on extremely short notice. The Marines have invested heavily in prepositioning strategic stockpiles of ammunition and equipment to prepare for contingent crises. But without getting the people to the stockpiles in time, in the worst situation, our prepositioned assets could be captured by an enemy and used against us. Therefore, the problem we addressed is one of great significance to our national defense. With our officer assignment system and a firm commitment to maintaining the WOSF and WASR data bases, the Marine Corps is ready to mobilize its officers quickly in war.

Appendix

Guidelines for Assignment Eligibility and Cost

Our mobilization system uses the following Marine Corps policies and preferences to decide whether an assignment arc should exist between particular officer-billet pairs and to decide how much existing arcs should cost. A nonretired officer who matches a billet perfectly with respect to grade, MOS, MCC, sex, and limited-duty status costs zero to assign. All other allowable assignments have positive cost.

- Active-duty officers are preferred to reserve officers for some high-priority billets
- Active-duty and reserve officers are preferred to retired officers in high-priority billets and, to a lesser extent, in other billets
- Females and limited-duty officers can never be assigned to billets from which they are restricted
- Grade substitution is most undesirable in high-priority billets (with the exception of some warrant officers who can fill lieutenant billets)
- Grade substitutions are permissible in medium-and low-priority regular billets under the following guidelines. These general guidelines are ignored, however, if specific guidelines are given for an individual MOS
    - Any officer can be assigned a billet that is one grade above his or her grade
    - Active-duty aviation officers, reserve officers, and retired officers can be assigned billets that are one grade below their grades
    - A retired officer can be assigned a billet this is two grades below
- Grade substitutions are permissible in low-priority generalized billets under the preceding guidelines
- Grade substitutions are prohibited when MOS substitutions take place
- In technical billets, MOS substitutions are worse than grade substitutions. In nontechnical billets, the reverse is true
- It is preferable to assign an officer to a billet requiring his or her PMOS rather than one of his or her AMOSs
- MOS substitution is permissible only for certain specified MOS pairs
- Billets in certain specified MCCs that are involved in the earliest mobilization actions have the highest priority
- Some reserve officers carry "hip-pocket orders" to report to specific MCCs in case of emergency. These officers should be assigned billets in the specified MCC
- High-priority billets should not be assigned to officers more than a specified number of miles away. Medium-priority billets have a similar, but less stringent, restriction

- Officers who are enrolled in the early weeks of certain basic MOS schools should not be given mobilization assignments. (They are screened out in the WOSF input step.)
- Retired officers cannot be used unless they retired less than a specified number of years ago. (This policy is also enforced through screening the WOSF on input.)

Several of these guidelines required specification of policy parameters. Our mobilization system stores default values in a small file that the user can edit at any time.

Brigadier General John J. Sheehan, USMC Director, Personnel Management Division Headquarters, US Marine Corps, writes:

> Tests have revealed that this system far exceeds the capabilities of our previous one. As a result of the increased capability, we now have the ability to quickly and efficiently determine appropriate placement of qualified officers to wartime billets. This means that during an all-out mobilization, our new system could provide the edge we need. As a result, it could effect the saving of lives and increase our probability of winning the war.
>
> Constraints of time, distance, and individual billet requirements could not be handled very well prior to the development of the officer mobilization assignment model. Since these constraints were primary considerations at every point of the development of our new system, we now have a tool that makes something, which used to be impossible, almost easy.
>
> We intend to use the knowledge that has been gained through this development to enhance our peace-time capabilities. We must ensure that our peace-time functions and ways of doing business do not interfere with or hamper any transition to wartime functions.
>
> If we must mobilize, we will be ready.

References

Bausch, D. and Brown, G. 1988. NDP FORTRAN and Phar Lap Tools: A PC environment for large-scale programming. *OR/MS Today*, (June).

Bradley, G., Brown, G., and Graves, G. 1977. Design and implementation of large-scale primal transshipment algorithms. *Management Science*, 24(1), 1–34.

Decision Systems Associates, Inc. 1983. *Officer Staffing Goal Model User's Guide: Enhanced OSGM*. Rockville, MD.

Gass, S., Collins, R. W., Meinhardt, C. W., Lemon, D. M., and Gillette, M. D. 1988. The army manpower long-range planning system. *Operations Research*, 36(1), 5–17.

Grinold, R. and Marshall, K. 1977. *Manpower Planning Models*. North-Holland, Amsterdam, The Netherlands.

Jarvis, J. and Shier, D. 1988. *User's Guide to NETSOLVE, URI*. Six Mile, SC.

Klingman, D., Mead, M., and Phillips, N. 1984. Network optimization models for military manpower planning. In *Operational Research '84*, J. P. Brans (ed). Elservier Science Publishers, Amsterdam, The Netherlands, 786–800.

Klingman, D. and Phillips, N. 1984. Topological and computational aspects of preemptive multicriteria military personnel assignment problems. *Management Science*, 30(11). 1362–1375.

Liang, T. and Buclatin, B. 1988. Improving the utilization of training resources through optimal personnel assignment in the US Navy. *European Journal of Operations Research*, 33(2), 183–190.

Liang, T. and Thompson, T. 1987. A large-scale personnel assignment model for the navy. *Decision Sciences*, 18(2), 234–249.

MicroWay, Inc. 1988. *NDP FORTRAN-386 User's Guide and Reference Manual*. Kingston, MA.

Rapp, S. 1987. Design and implementation of a network optimizer for officer assignment during mobilization. MS Thesis, Naval Postgraduate School, Monterey, CA.

SAS Institute, Inc. 1985. *SAS User's Guide: Basics, Version 5*. Cary, NC.

Silicon Valley Software. 1990. *SVS FORTRAN 77 User's Guide and Reference Manual*. San Mateo, CA.

18

The Deployment Scheduling Analysis Tool (DSAT)*

Thom J. Hodgson
North Carolina State University

Barbra Melendez
United States Military Academy

Kristin A. Thoney
North Carolina State University

Tim E. Trainor
United States Military Academy

18.1 Introduction

When asked about the secret of his many successful military operations, Confederate Cavalry General Nathan Bedford Forrest answered that victory often goes to "…whoever gets there firstest with the mostest…" The simple logic of this statement underscores the fact that military victories are critically dependent upon the military transportation system. The military must be able to effectively plan, analyze, schedule and execute force deployment operations. Many deployments are short-notice contingency operations that require rapid planning and scheduling.

Currently, deployment planners have few software tools available to assist them. One tool, the Joint Flow and Analysis System for Transportation (JFAST) [8], is a simulation model that provides detailed estimates for strategic force movement by air and sealift assets. JFAST requires a rigid input structure involving considerable time to prepare, and is a good tool for sensitivity analysis. Another tool, the Deployment Analysis Network Tool Enhanced (DANTE) [6], is a valuable tool for deployment planning sensitivity analysis. DANTE is extremely fast to use, but lacks the detail necessary to construct a useable deployment schedule.

The need for computational speed and detailed estimates drove the development of a PC-based tool for deployment scheduling and analysis. Illingworth [5], then Trainor [13] initially developed the Deployment Scheduling Analysis Tool (DSAT). DSAT provides a flexible and rapid means for analysis of

* This article was first published by Elsevier: Reprinted with permission. Hodgson, T.J., B. Melendez, K.A. Thoney and T.E Trainor. 2004. The deployment scheduling analysis tool (DSAT), In *Defense Transportation: Algorithms, Models, and Applications for the 21st Century*, ed. R.T. Brigantic and J.M. Mahan. Elsevier, San Diego, pp. 905–924.

deployments. It is quick and flexible enough to be used for sensitivity analysis of various aspects of the deployment scenario. At the same time, enough information concerning the deploying forces is retained that detailed equipment schedules (what equipment moves over which routes at which time) can also be analyzed. A visual basic (VB) graphic user interface (GUI) facilitates modeling of deployment scenarios and viewing of key outputs for analysis. User inputs, coupled with data drawn from military databases, provide the information needed to schedule the movement of equipment and cargo on transportation assets and deployment routes. A specially tailored job shop scheduling procedure [4,11,12] performs the scheduling. The user can invoke a heuristic procedure to move transportation assets between routes to affect improved schedules. The overall logic used in DSAT is represented in Figure 18.1 and detailed in the rest of this chapter. Section 18.2 introduces DSAT and presents the steps used to model a problem. Section 18.3 describes the heuristics used to create a deployment schedule and to determine an alternate transportation asset to route allocation. Section 18.4 discusses the various reports provided and the options available for sensitivity analysis. Section 18.5 presents conclusions.

18.2 Modeling a Deployment Problem with DSAT

The intended use of DSAT requires ease of use in developing deployment scenarios and speed in calculating the corresponding "optimal" schedule. While DSAT models the deployment problem in a good deal of fidelity, it should not be used as a final scheduling tool for an actual deployment. Rather DSAT enables planners to quickly develop and analyze schedules for possible deployment courses of action in order to reduce the set of viable alternatives for scheduling using more detailed, less flexible, deployment simulations. DSAT also enables planners to rapidly perform sensitivity analysis to determine the impact on deployment schedules by changing key parameters. To meet the intended use, some assumptions are modeled in DSAT that do not necessarily reflect the many uncertainties that affect a real world deployment. Specifically, DSAT assumes transportation assets are operational 24 hours a day, although standard ground load, unload and maintenance times are factored into asset availability. Port facility restrictions are known in advance and remain fixed during the deployment, as do asset-operating characteristics. While not explicitly capturing the many uncertainties to which a real world deployment is subject, DSAT retains the level of fidelity necessary to create an "optimal" equipment and asset schedule yet has the speed and flexibility needed to conduct deployment sensitivity analysis.

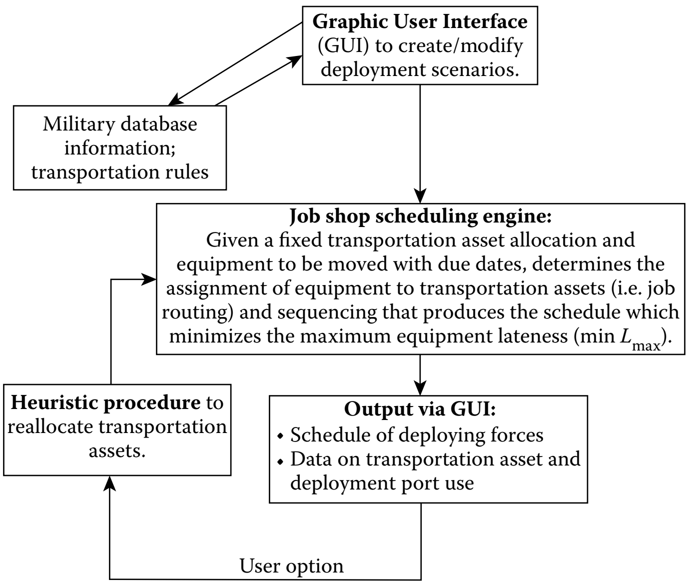


FIGURE 18.1 Overall DSAT methodology.

A deployment problem is built by retrieving and storing information in Microsoft Access© database tables. The DSAT database contains information for several military units and the characteristics of their equipment, global position and capacity data on ports throughout the world, and characteristics of aircraft, railcars and ships used by the military. Using structured query language (SQL), the appropriate information is retrieved from these tables to formulate the problem. Deployment information is grouped into four areas:

- Deployment units and equipment
- Ports through which deployment units are routed
- Deployment time windows
- Characteristics of transportation assets

DSAT builds the problem in this sequence as the user navigates through a series of GUI screens.

18.2.1 Determining the Movement Requirement

To build the deployment force, the user first selects a major unit from an interface screen as depicted in Figure 18.2. Figures 18.2 through 18.10 show the various DSAT GUI screens users navigate through to model a deployment scenario. When the user selects a subordinate unit for deployment by checking the box next to its name, the unit's equipment is added to the deployment equipment list by retrieving it from the DSAT database. The user navigates through the interface screens of Figures 18.2 and 18.3 for each major unit desired, and thus builds the deploying equipment list.

18.2.2 Selecting the Deployment Ports

The set of ports used is another major component of the deployment network. The chosen ports determine the deployment route distances used to calculate the travel times for the transportation assets. Each port is uniquely identified in the DSAT database by a four-character code termed a geo-location code. The code identifies the port in table containing geographical position, distance and any known facility data.

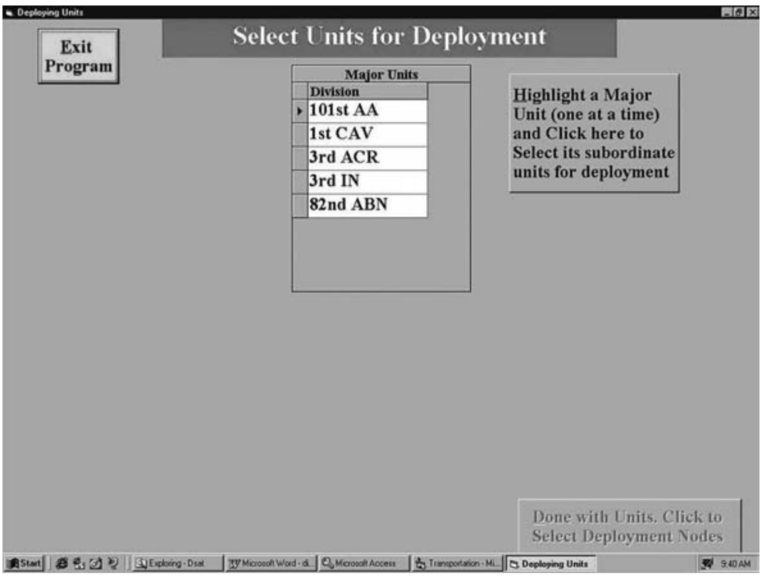


FIGURE 18.2 Major units.

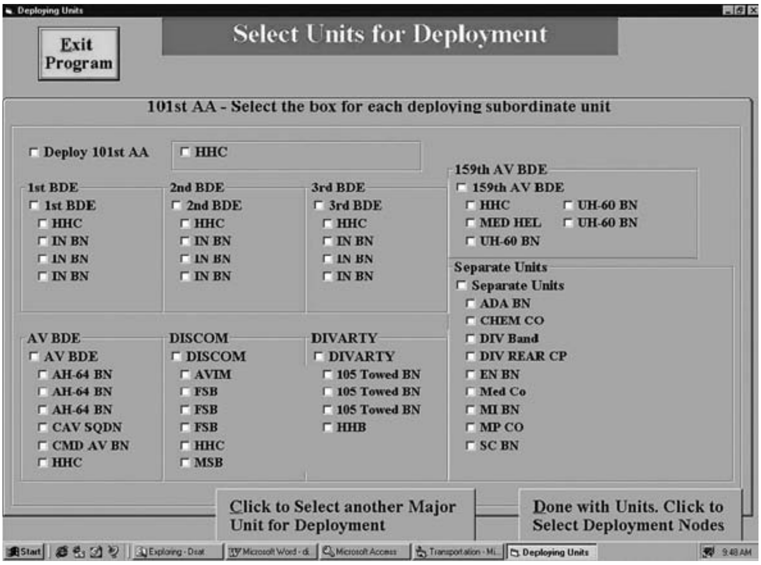


FIGURE 18.3 Selecting subordinate units for deployment.

All deployment equipment leaves its home base (or current physical location) by air or rail transportation. In the DSAT database, each unit is linked to a habitual airport of embarkation (APOE), seaport embarkation (SPOE) and rail depot from which equipment travels to the SPOE. The user verifies and changes, if necessary, these deployment ports using the interface screen depicted in Figure 18.4. The intent is to allow the user maximum flexibility in developing realistic deployment scenarios.

Based on the desired deployment scenario, the user selects any intermediate staging bases (ISBs) desired and the final destination ports. This selection process is shown in Figure 18.5. Once the user selects the countries and the various hubs (ports) to be used, the geo-location codes are used to retrieve distance data.

If no distance data exists in the tables for a chosen set of ports, position data is used to calculate the great circle distance between two selected ports. DSAT uses the Haversine formula (see Sinnott [10] and Williams [14]) to calculate the great circle distance, d, between two points with coordinates $\{lat1,lon1\}$ and $\{lat2,lon2\}$. The equation is a follows:

$$d = 2 * R * \arcsin\left(\sqrt{\sin^2(lat1 - lat2/2) + \cos(lat1) * \cos(lat2) * \sin^2(lon1 - lon2/2)}\right) \tag{18.1}$$

where R is the earth's radius. The great circle distance calculator used in DSAT is a modified version of a function developed by North [9], with formulae accredited to Chamberlain [11] and Greer [3]. This is an accurate means for calculating the distance traveled by aircraft between airports, but might not accurately reflect actual travel distances for rail and sea routes. The user has an option to overwrite the great circle distances generated by the program.

Ports have facility characteristics that also affect the deployment problem. Airports have a limited number of runways that limit throughput capacity. The available space for aircraft, fueling capacity and loading and unloading capability are also limited. These factors are modeled by the maximum-on-ground (MOG) parameter that constrains the number of aircraft on the ground at an airport at a given time. The number and type of berths at a seaport, along with the type of material handling equipment available, determine the number and type of ships that can be loading or unloading. The size of rail depots and the type of loading ramps available dictate the number of cars that can be loading or unloading. These facility restrictions affect the amount of cargo capable of being processed through the port in a specific time period.

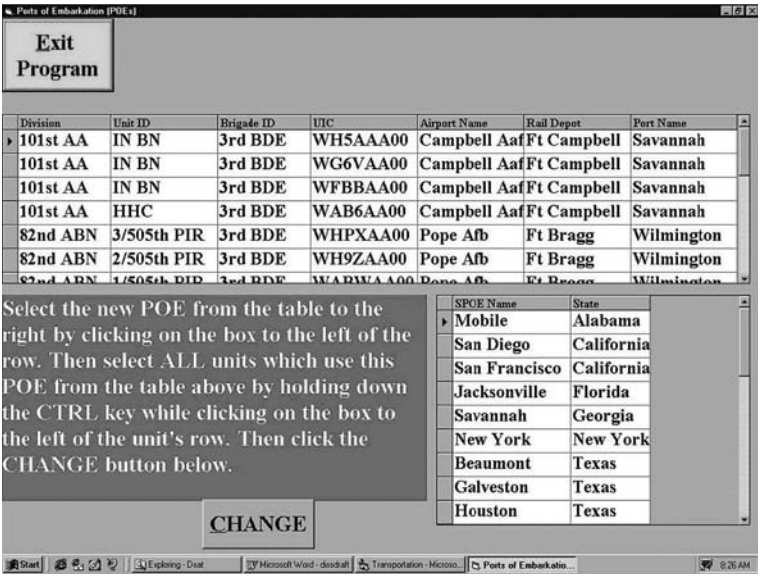


FIGURE 18.4 Verifying the ports of embarkation.

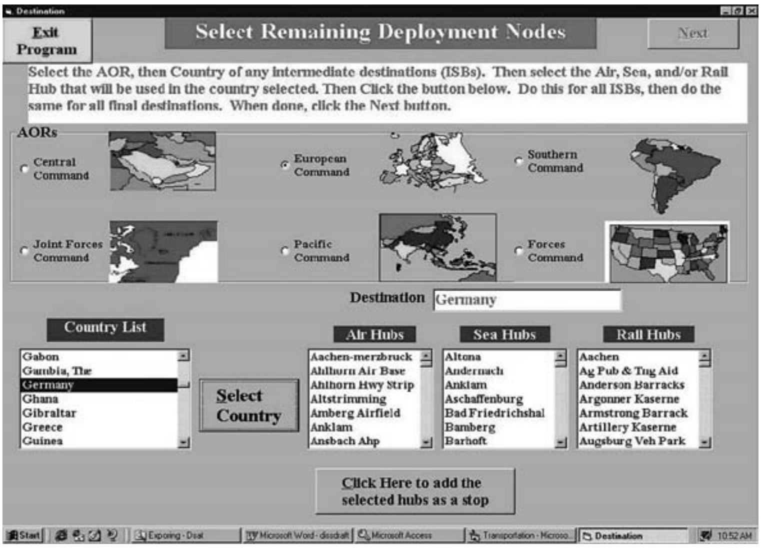


FIGURE 18.5 Selecting additional deployment ports.

DSAT explicitly incorporates the number of runways and MOG restrictions, and the number of berths available as simple numerical parameters. The DSAT database lacks airport MOG data since the USAF Air Mobility Command (AMC) closely guards this information, while berth information is available for many of the major seaports of the world. The user can update available information for seaports and airports using the interface depicted in Figure 18.6. DSAT does not verify that transportation assets chosen by the user match the capability available at a selected port. Transportation planners have the requisite information available to select assets that can be accommodated at the selected ports. DSAT does not account for facility restrictions in rail depots. While loading and unloading times for railcars are explicitly incorporated, the user must select a realistic number of cars that can efficiently operate between chosen rail depots.

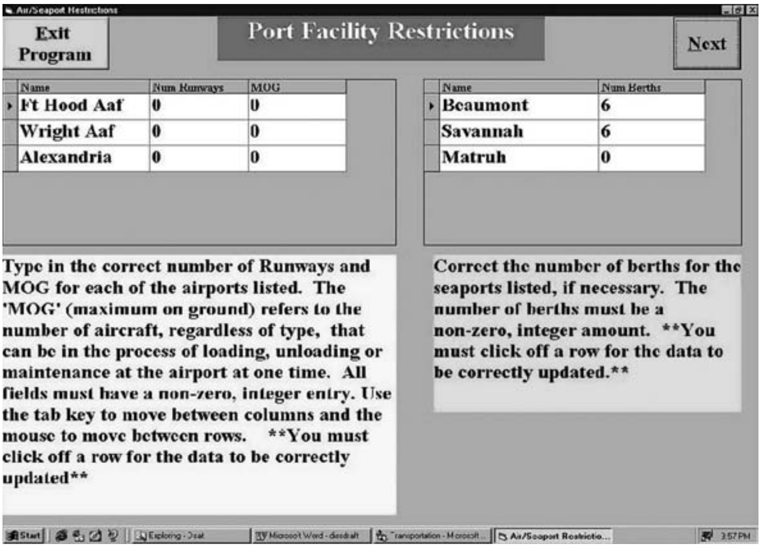


FIGURE 18.6 Entering port facility restrictions.

18.2.3 Determining the Deployment Time Window

The Unified Combatant Command determines the required delivery date (RDD) for each deploying unit based on the operational plan. Transportation planners are expected to backwards plan the deployment of units to meet these RDDs. Another time constraint is the unit available to load (ALD) date. That is the time equipment is available to start movement from its home base or current location.

DSAT accepts user-input for the ALD; this ALD applies to all deploying units. The user selects an RDD for each unit as shown in Figure 18.7. In scheduling terms, the ALD is the release date and the RDD translates to the due-date for each item of equipment in the unit. These release and due dates play a significant role in the scheduling methodology employed by DSAT.

18.2.4 Assigning Transportation Assets

The user assigns transportation assets to specific deployment routes in DSAT that are developed from the user selected ports. The military users various types of planes, trains and ships, each with their own set of operating characteristics. DSAT enables the users to assign multiple types of transportation assets to a route and vary their operating characteristics. As shown in Figure 18.8, the user simply follows the onscreen instructions to change any characteristic of interest.

The deployment routes are presented to the user in the screen displayed in Figure 18.9. The routes are segregated by air, rail and sea. Assets are assigned to each route by using a screen similar to that in Figure 18.10. This figure represents an example of assigning aircraft to a deployment route from Pope Air Force Base, NC, to the Canchones West Airfield in Chile. The route distance appears in the top right corner. The user may edit the distance data and follow the onscreen directions to assign assets to the routes. The transportation asset assignment is complete after repeating the process for each route.

18.2.5 Feasibility Checks of the Deployment Problem

Once the deployment problem is created, DSAT performs a set of feasibility checks, modifies the inputs as required, and builds a text input file used to schedule the deployment.

DSAT performs a check to insure the user has created a feasible deployment problem with equipment fit on the selected transportation assets being the primary focus. All military equipment will fit on the

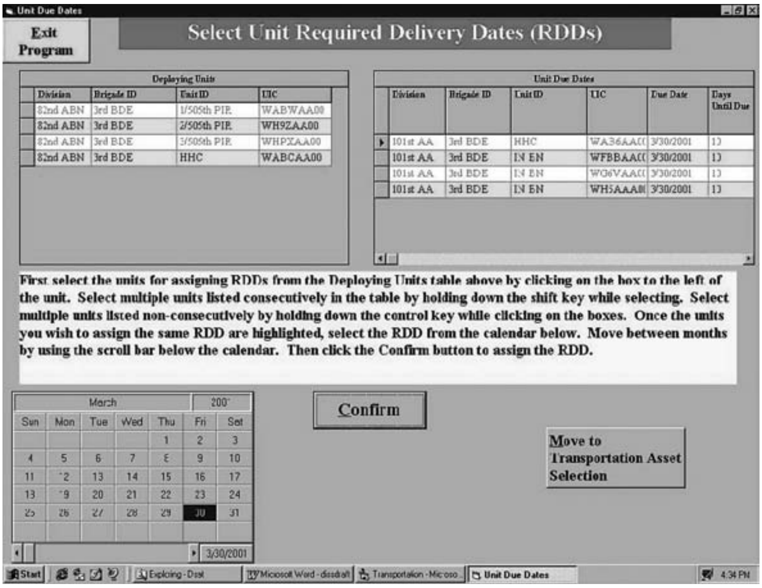


FIGURE 18.7 Assigning RDDs.

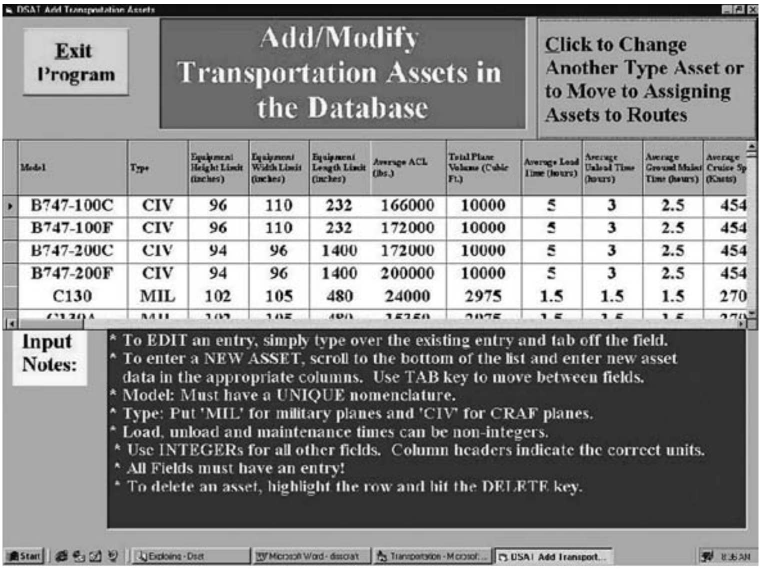


FIGURE 18.8 Modifying aircraft operating characteristics.

set of ships and railcars in the DSAT database. Certain items may need overhang clearance to travel on some railcars. DSAT generates a list of items needing overhang rail clearance as an output report.

The most crucial equipment fit checks for aircraft. Individual pieces of unit equipment are checked to see if they exceed the total weight and volume capacities of the planes used in the deployment asset fleet. Equipment dimensions are also compared to the dimensions of the aircraft-loading doors. From the checks, a list of feasible transportation assets on which each item of equipment can fit is passed to the scheduling procedure.

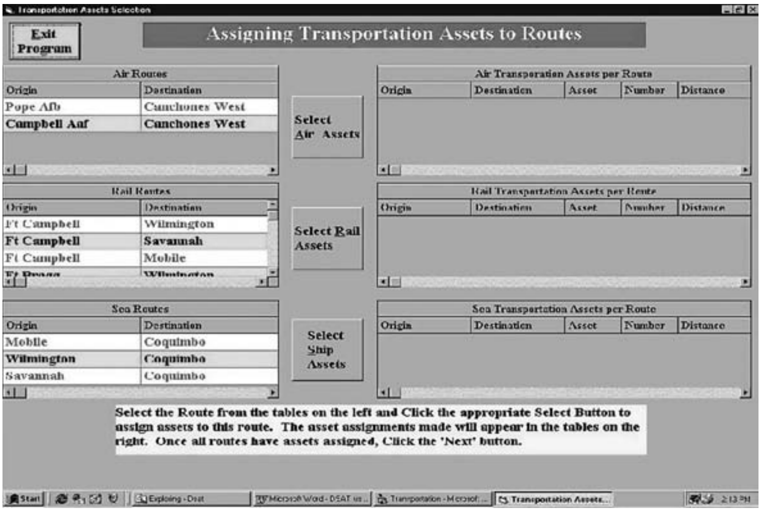


FIGURE 18.9 Assigning transportation assets to routes.

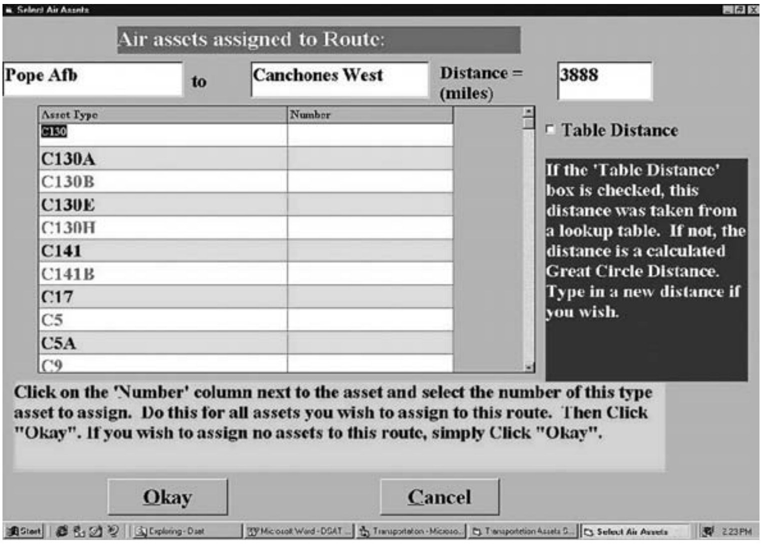


FIGURE 18.10 Example of assignment of aircraft to a route.

If the user selects no assets on which a certain deployment item of equipment will fit, DSAT provides a series of options to correct the problem. The user can view this list of equipment and choose to remove the items from the deploying equipment list or return to assigning transportation assets to correct the problem. DSAT prompts the user to the smallest-sized aircraft on which the item of equipment will fit. DSAT provides an output report of any items of equipment removed from the deployment.

18.3 DSAT Scheduling and Asset-route Reallocation Routine

The DSAT software package contains two key routines. The interface discussed above, is written in VB and allows the planner to develop scenarios and conduct sensitivity analysis. The first routine, the scheduler routine, is written in C++. It creates a (heuristically optimized) detailed equipment deployment schedule by allocating equipment to routes and then sequencing that equipment onto the transportation assets. The scheduler routine attempts to meet the RDD for all equipment. The second routine, the asset-route reallocation heuristic, is written in VB. It reallocates transportation assets to alternative routes in the deployment network to reduce the total time to close on the objective.

18.3.1 Military Application of the Tailored Job Shop Scheduler

DSAT takes the input information from the planner and generates a text file appropriate for the scheduler routine. The information passed includes the deployment routes, the facility restrictions of each port, the number and operating characteristics of each type transportation asset assigned to each route and unit equipment data. The equipment data includes the delivery time window, start location, dimensions and a vector listing the types of transportation assets on which the item fits.

Hodgson et al. [4] demonstrated that a conceptually simple simulation-based procedure (first proposed by Lawrence and Morton 1986) is both effective and efficient in providing optimal or near optimal schedules for minimizing the maximum lateness (L_{max}) in large job shops. Thoney [11] and Thoney et al. [12] reports on the development of the military scheduling procedure based on scheduling networks of factories; this became DSAT's scheduling routine. Each item of unit equipment being deployed represents one "real" job to be scheduled (fake jobs are created as "space fillers" for the various transportation assets). Airports and seaports are "factories" through which a job must process. Transportation assets, "vehicles" operating on routes, are batch processors on which jobs move between "factories". Different asset types operating between the same two ports are considered different asset-routes. The procedure assigns the equipment to transportation asset-routes and then sequences the equipment onto the assets to minimize the worst case difference between the equipment arrival at destination and its required delivery date or RDD (i.e. schedule to minimize Lmax).

18.3.2 Scheduler Routine Heuristics

The scheduler routine incorporates two unique heuristics. These heuristics work together to allocate the jobs to the paths (sequence of asset-routes from origin to final destination) in an attempt to minimize the maximum lateness over all jobs. Jobs are presorted by due date. To minimize maximum lateness, the job to path allocation scheme attempts to schedule the jobs' arrival at the final destination by their due date. The initial allocation heuristic constructs the path that each job will travel over based on the average travel time to complete the path. The jobs are initially assigned to the path that is anticipated to be the fastest.

The reallocation heuristic makes use of information gained through simulating the previous allocation. In the simulation, as vehicles arrive and are unloaded at the base(s), arrival times are recorded. Jobs usually do not arrive in exact due date order. Therefore, based on this information, jobs then are reallocated to paths according to actual arrival at the final destination. This process is repeated a set number of times. The best schedule is saved.

The amount of real time needed to create the equipment schedule varies according to the size and complexity of the deployment scenario. However, DSAT's scheduling engine creates the deployment schedule in minutes. For example, a two-division sized force (17,942 pieces of equipment) deployed by rail, sea and air was schedule in slightly over 2 minutes on a 700-megahertz PC.

18.3.3 Initial Allocation Heuristic

The initial allocation heuristic builds each job's path based on the structure of the deployment network. To build the paths, multiple passes through the network are required. During the first pass, a backward traverse of the network, the average "remaining time" from each base to the final destination is computed using all possible paths. During the next passes of the network, forward traverses, each job at the current base, base i, is allocated to a vehicle (deployment transportation asset) that has the earliest anticipated final destination arrival time and can accommodate the job.

At each base, the jobs are sorted in due date order. Once the job reaches the next destination in the path, it is inserted into the list of jobs at the base. Job scheduling at base i can not begin until all jobs at bases preceding base i have been scheduled. Then all jobs scheduled to move through base i are on base i's job list, and job allocation at that base can now begin. The anticipated final destination arrival time of jobs allocated on a vehicle is the vehicle's "job time" plus the average remaining time to reach the final destination base from that location. The "job time" is the amount of time needed to load and unload the vehicle in addition to the one-way travel time needed to reach the vehicle's assigned destination. The earliest that subsequent trips for each vehicle can start is after the vehicle returns to it's origin base.

In the deployment network, multiple asset types can operate between the same two parts. Each asset type operating on such a route is called a center; each center has a number of vehicles operating on that route. The following additional definitions are necessary for both the initial and the reallocation heuristics:

N, number of jobs
U_i, area of job i
W_i, weight of job i
V_i, volume of job i
Q_i, current path (in terms of centers) of job i
B_{il}, Boolean value indicating if job i can be placed on center l
M, number of centers
m_j, number of vehicles within center j
TU_j, total area capacity of each vehicle in center j if j is a ship, otherwise $Tu_j=-1$
TW_j, total weight capacity of each vehicle in center j
TV_j, total volume capacity of each vehicle in center j, if j is a not a ship, otherwise $TV_j=-1$
K_j, vehicle type of center j
J_j, time job spends on a vehicle from center j (load time+one way travel time+unload time)
H_j, time to complete return trip of vehicle from center j (maintenance time+one way travel time+maintenance time)
T, trip number of a vehicle
R_b, average remaining time from base b to the final destination base
A_{jm}, the anticipated final destination arrival time if a job is schedule on the mthtrip of a vehicle in center j
D_j, destination base of center j
$C(b)$, set of centers originating at base b
$L(b)$, set of jobs currently located at base b
$F(b)$, set of fake jobs currently located at base b

Note the conditional definitions of TU_j and TV_j. The scheduler routine maintains information concerning the weight and area capacities for all ships throughout the simulation. The scheduler also maintains information concerning the weight and volume capacities for all other types of transportation assets (trains and planes).

The first trip of a vehicle in center j is anticipated to arrive at the final destination at time $A_{j1}=J_j+R_{Dj}$. The second trip of a vehicle in center j is anticipated to arrive at the final destination at time $A_{j2}=J_j+H_j+J_j+R_{Dj}$. Consequently, the Tth trip of a vehicle in center j is expected to arrive at time $A_{jT}=T(J_j)+(T-1)H_j+R_{Dj}$ The average remaining time from base b to the final destination base is 0 if b is the final destination base, otherwise it is calculated as:

$$R_b=\frac{\sum_{x\in C(b)} m_x\,(J_x+R_{Dj})}{\sum_{x\in C(b)} m_x}$$

The algorithm for the initial allocation heuristic has three parts; each part is carried out for every base. The first part involves creating a list of vehicle nodes sorted in increasing order of anticipated arrival time at the final destination. There is one vehicle node list created for each base. Let node (s, t, u, v, w) be a vehicle node where s is the anticipated arrival time at the final destination, t is the center number, u is the available area capacity, v is the available volume capacity, and w is the available weight capacity. For each vehicle, one node is placed on the list representing its first arrival at the final destination. The first part of the heuristic (the backward traverse) is detailed as follows:

For each base b:
  For each $j \in C(b)$: (for each center operating at base b)
    Step 1. Calculate A_{j1}
    Step 2. Set vehicle k within center j equal to 1
    Step 3. Insert node (A_{j1}, j, TU_j, TV_j, TW_j) on the vehicle node list
    Step 4. if $k < m_j$, set k to $k+1$, go to Step 3
  Next center j
Next base b

The completion of this part of the initial allocation heuristic results in a vehicle node list at each base sorted in increasing order of anticipated arrival times at the final destination. A node for each vehicle in each center is placed on this list. Each entry details the cargo's first anticipated arrival time at the final destination, the center number and the vehicle's available capacities. This provides a structure for initially assigning jobs to vehicles based on the job due dates.

The second part of the initial allocation heuristic (a forward pass) allocates the real jobs on the vehicles and adds subsequent vehicle trips on the vehicle node list as necessary. This part's algorithm is as follows:

For each base b: (starting at origin base)
  For each $i \in \{L(b) - F(b)\}$: (for each real job at base b)
    Step 1. Scan vehicle node list from beginning; find the first node (s, t, u, v, w) simultaneously satisfying the following conditions:
      $U_i \leq u$ (if $u \neq -1$) (job area fits into remaining vehicle area capacity)
      $V_i \leq v$ (if $v \neq -1$) (job volume fits into remaining vehicle volume capacity)
      $W_i \leq w$ (job weight fits into remaining vehicle weight capacity)
      B_{it} = True (job i fits on center t)
      If no such node exists, go to the next real job
    Step 2. If $u = TU_j$, and $v = TV_j$, and $w = TW$j, insert node ($s + J_t + H_t$, t, TU_t, TV_t, TW_t) on list (create next trip)
    Step 3. For the node found in Step 1, Let $u = u - U_i$ (if $u \neq -1$), $v = v - V_i$ (if $v \neq -1$) and $w = w - W_i$. (decrement remaining capacities)
    Step 4. Let $Q_i = Q_i + \{t\}$ (augment current path with center t)
    Step 5. Set $L(D_t) = L(D_t) + \{t\}$ (insert job on job list at next destination)
  Next i
Next b

This part of the initial allocation heuristic assigns each job to move on a specific center vehicle. If the job fits into the remaining volume (or area) and weight capacities of the vehicle and is compatible with the center, then that job is allocated to that vehicle. If the vehicle is empty, a subsequent trip for that vehicle is added to the vehicle node list (Step 2), while the remaining capacities of the vehicle for its current trip are decremented (Step 3). The path of the job is updated to indicate job travel on that center (Step 4). Finally, the job is added to the job list at the center's destination base (Step 5). Any job that does not fit into any vehicle originating at that base is essentially removed from that base's job list. These jobs are assigned a path at the end of the initial allocation heuristic. This results in each job having a specific path built for it. The path is expressed as the sequence of centers the job travels on to get from its origin to its final destination.

In creating the initial job to path allocation, the initial allocation heuristic could over or under utilize a particular path. This situation, in turn, reduces the options available to the reallocation procedure that shifts jobs to alternate paths. To alleviate this problem, *fake jobs* are created for each path. These fake jobs have very low priority (i.e. their due dates are much later than real job due dates). Allocating these fake jobs to paths creates additional path arrival times at the final destination. Thus, the reallocation procedure has more options available for all paths in considering job rerouting. Consequently, the third part of the initial allocation heuristic creates the fake jobs and their paths. Once again, the bases are treated individually, traversing in a forward manner. On fake job, i, is initially placed at each origin base with $U_i = V_i = W_i = \infty$.

The procedure is as follows:

For each base b:
  For each $i \in F(b)$: (for each fake job)
    Step 1. Let $temp = 0$
    For each $j \in C(b)$: (for each center)
      Step 2. Set $temp = temp + 1$
      If $temp \neq |C(b)|$ (if $temp \neq$ number of centers at base)
        Step 3. Create a fake job, d
        Step 4. *Let* Let $Q_d = Q_i + \{j\}$
        Step 5. If $TU_j \neq -1$, let $U_d = \min\{U_i, TU_j\}$,
          If $TV_j \neq -1$, let $V_d = \min\{V_i, TV\text{j}\}$,
          Let $W_d = \min\{W_i, TW_j\}$
        Step 6. Let $L(D_j) = L(D_j) + \{d\}$
      Else
        Step 7. Let $Q_i = Q_i + \{j\}$
        Step 8. If $TU_j \neq -1$, let $U_i = \min\{U_i, TU_j\}$,
          If $TV_j \neq -1$, let $V_i = \min\{V_i, TV_j\}$,
          Let $W_i = \min\{W_i, TW_j\}$
        Step 9. Let $L(D_j) = L(D_j) + \{i\}$
      End if
    Next j
  Next i
Next b

When this procedure is complete, each path has one fake job assigned to it that has the maximum possible weight and volume a job can have if it is allocated to that path. The assignment of fake jobs to paths is done in both Steps 3 through 6 and Steps 7 through 9. Steps 3 through 6 basically clone each fake job that is currently at the base. The cloned jobs are assigned to the same path taken by the current fake job in arriving at the current base. Then each of the cloned jobs is designated for transport by a unique center at the current base. Steps 7 through 9 assign the last fake job transport by the remaining center.

After this procedure is complete, the smallest possible area, volume and weight over all of the paths are found. Each fake job is then cloned 50 times and the clones' area, volume and weight are set equal to these smallest values. The purpose for having 50 smaller jobs and one large job on each path is to help ensure that at least one extra trip is created for each vehicle on the path and that remaining capacity on partially filled "final" trips of vehicles in all centers are completely filled with fake jobs. The purpose of this is to produce additional arrival times at the destination for each path; this provides the reallocation procedure more options in rescheduling jobs. In forming the reallocation list, blocks are created from aggregating job capacity regardless if a job is real or fake. In contrast, only the real jobs are rescheduled into the blocks. The fake jobs have a fixed path.

The final part of this heuristic creates paths for real jobs that have not yet been allocated. Each remaining real job is compared with the fake jobs (each fake job has a fixed path) until a compatible path is found. The real job is then allocated to that path.

18.3.4 The Reallocation Heuristic

Once the jobs are scheduled by the initial allocation heuristic, that job allocation is simulated by the scheduler routine. Information is gathered on the actual arrival times of each job to the final destination. Each time a job arrives at the final destination, its area, volume and weight is added to the aggregate area, volume and weight of jobs that already arrived by way of the same path at the same time. The arrival time information is thus used to create *blocks of used capacity* that arrive at the final destination at given times. The reallocation heuristic reschedules the jobs (sorted in order of increasing due dates) to these blocks of used capacity sorted in increasing order of final destination arrival time.

The reallocation heuristic has two parts. Each part is performed once. The first part involves developing a list of used capacity blocks during the simulation sorted in increasing order of final destination arrival times. The second part of the heuristic entails the actual reallocation of the jobs to the vehicles. The reallocation heuristic uses many of the terms defined for the initial allocation heuristic, but several additional terms are defined next. Let block (s, t, u, v, w) be a used-capacity block where s is the final destination arrival time, t is the tth arrival at time s,u is the current used area capacity, v is the current used volume capacity and w is the current used weight capacity for the block. The following additional definitions are also necessary:

P_{qr}, current path (in terms of centers) of the r^{th} block arriving at time q
$E(s)$, set of jobs arriving at the final destination at time s
NB_s, total number of blocks arrived at time s

The first part of this heuristic records the final destination arrival times and aggregates certain used capacities into blocks of used-capacities. It is detailed as follows:

Step 1. Let x be the first time in which at least one job arrives at the final destination.
Step 2. Set NB_x equal to 1 (initialize total number of blocks arriving at time x)
Step 3. For each $i \in E(x)$: (for each job that arrives at the final destination at time x)
   a. Search list of all blocks (s, t, u, v, w), where $s = x$ to see if $P_{st} = Q_i$ for any t (find a block arriving at time x with the same path as job i)
   b. If a block exists at some t (number of arrival, not time), let $u = u + U_i$ (if $u \neq -1$), let $v = v + Vi$ (if $v \neq -1$) and let $w = w + W_i$. Otherwise increment NB_x by 1 and add block (x, NB_x, U_i, V_i, W_i) to the list.
   Next i
Step 4. If all jobs are not at the final destination, let x be the next time at least one job arrives at the final destination and go to Step 2.

The result of this first part of the reallocation heuristic is a list of used capacity blocks sorted in increasing arrival time at the final destination. Each used capacity block is then inflated by 20%; i.e. the used area, volume and weight are increased by 20%. The purpose of this inflation is to reduce the amount of unused capacity on the blocks' path in preparation for the eventual job rescheduling.

The second part of the algorithm performs the rescheduling of the real jobs (sorted in increasing due date order) onto the identified used-capacity blocks. If a real job cannot be rescheduled on its current path. The fake jobs always keep their initial path. The second part of the heuristic is as follows:

Step 1. Set the job, i, equal to 1 (current job to be scheduled).
Step 2. Search the list of blocks to find the first block (s, t, u, v, w) that simultaneously satisfies the following conditions:
Block has same origin base as job i
$U_i \leq u$ (if $u \neq -1$) (job fits into remaining used area capacity)
$V_i \leq v$ (if $u \neq -1$) (job fits into remaining used volume capacity)
$W_i \leq w$ (job fits into remaining used weight capacity)
B_{ik}=True $\forall k$ in P_{st} (job fits onto all centers in path)
Step 3. If no such block exists, go to Step 6, otherwise, go to Step 4.
Step 4. If $u \neq -1$, let $u = u - U_i$. If $v \neq -1$, let $v = v - V_i$. Let $w - W_i$ (decrement remaining block capacities)
Step 5. Let $Q_i = P_{st}$ (redesignate path for job i).
Step 6. If $i < N$, increment i by 1 and go to Step 2.

This results in jobs being reassigned to paths based on previously demonstrated performance, i.e. final destination arrival time established by the previous deployment simulation. Note, during the simulation, vehicles are loaded and unloaded at bases. No vehicle is loaded beyond its user-defined capacities for weight and volume or area.

18.3.5 Redefining the Asset-route Allocation

The assignment of jobs (equipment) to assets produced by the scheduling routine is limited by the fixed allocation of transportation assets to routes used in DSAT. The transportation asset-route allocation, input by the planner, may not necessarily by the best possible. The user has the option to activate a heuristic procedure that reallocates assets seeking to find an improved schedule.

The basic concept of this procedure is to shift underutilized transportation assets to the route(s) on which the latest pieces of equipment move. By adding capacity to these "late job" route(s), congestion and possible queuing experienced by the equipment may be reduced, leading to an improved schedule in terms of better satisfying the RDD's (i.e. a lower $L_{\max}$ value). It can be time-consuming (approximately 5 minutes for a two brigade sized deploying force) due to the supporting calculations needed; this heuristic is limited to a maximum of three iterations. The logic of this heuristic is shown in the flowchart depicted in Figure 18.11.

18.4 Output and Sensitivity Analysis

The interface screen displayed in Figure 18.12 provides information on deploying unit closure at the final destination and transportation asset use on each route. The top part of the screen provides unit closure results. The bottom table provides data concerning transportation asset-route utilization. The top right portion of the screen shows the aggregate closure information. In this example, the deployment of 40 units with a total of 9762 items of equipment was completed in C+47 days (C day is the ALD) with 26 units arriving beyond their specified RDD. The top table provides a row for each unit with its actual closure date. If a unit did not close by its RDD, the information is presented in red.

The user can choose to look at detailed results of the current schedule or can conduct sensitivity analysis by modifying the scenario and rescheduling the deployment. To analyze the results of the current schedule, the user can look at various graphs depicting the weight or volume utilization history for a specific asset-route couple or view various reports concerning asset schedules, unit closure profiles, etc. To perform sensitivity analysis, the user can modify the scenario by changing unit due dates or the transportation asset-route allocation and reschedule the deployment to determine the impact of the change(s).

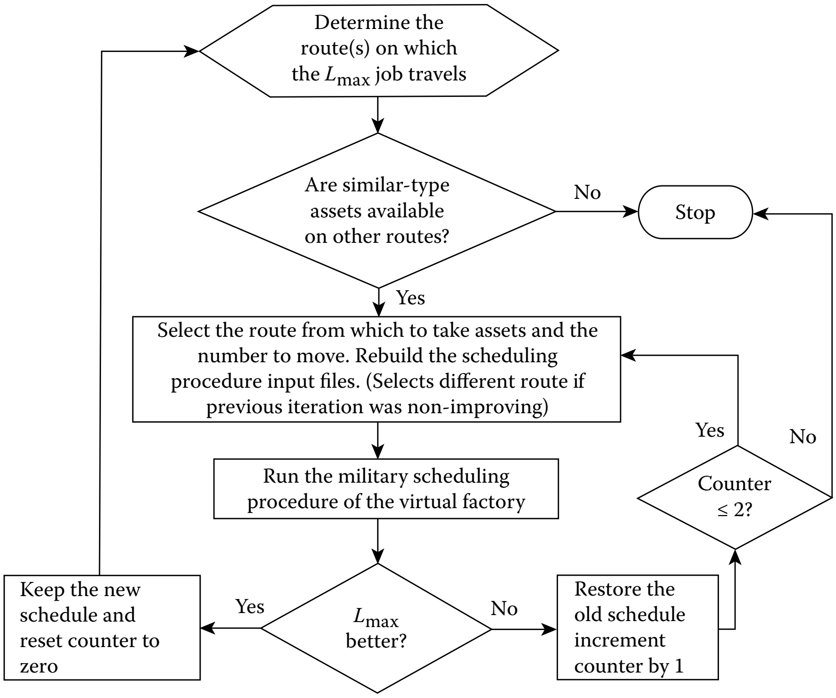


FIGURE 18.11 Asset-route reallocation heuristic.

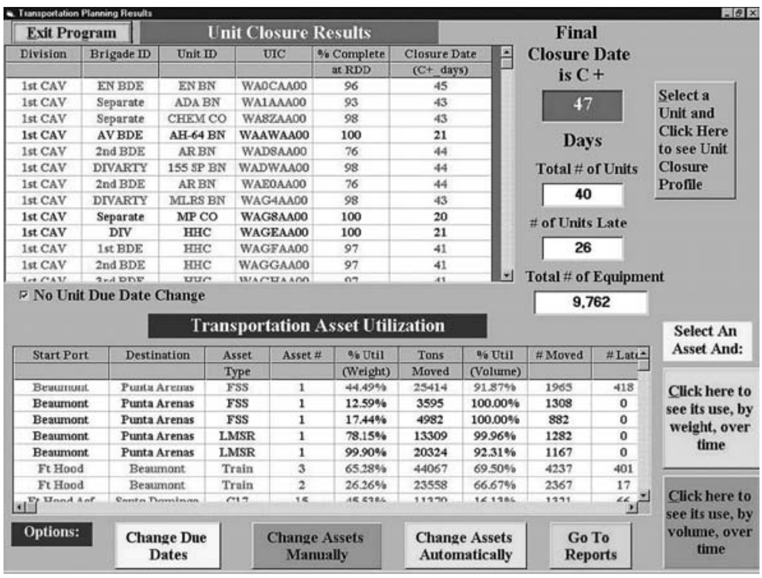


FIGURE 18.12 Output of DSAT scheduling results.

DSAT graphically represents the arrival of units to the final destination over time. A unit's closure profile can be generated for any unit selected from the top table in Figure 18.12. An example is provided in Figure 18.13. The graph is calculated simply by summing the weight of unit equipment delivered to

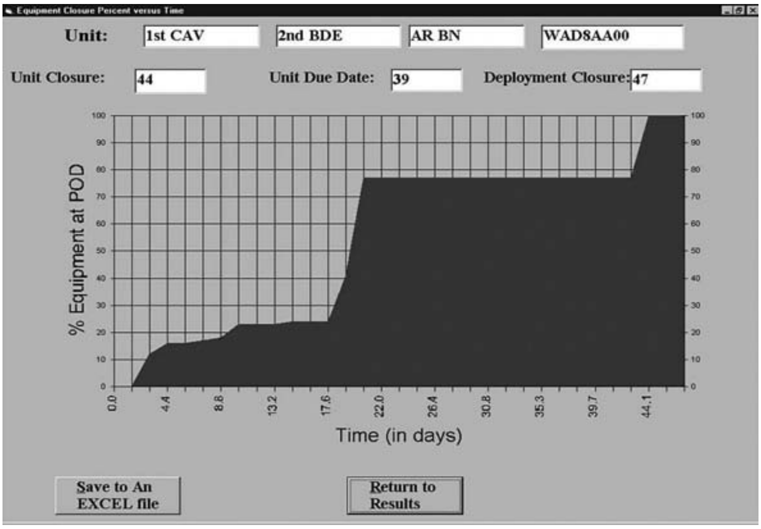


FIGURE 18.13 Example unit closure graph.

the port of debarkation (POD) at specific time intervals. The graph displays the timing of unit equipment arrival to the POD along with pertinent unit and closure information. All graphs may be saved for future use in a Microsoft Excel© file.

18.4.1 Utilization of Transportation Assets

The bottom table in Figure 18.12 provides key information on the use of transportation assets. Each row represents the assets assigned to a particular route. For example, row one of the table represents one fast sealift ship moving cargo from Beaumont, TX, to Punta Arenas, Chile. Each ship sailing between the same origin and destination represents one route each ship tends to have unique operating characteristics. Figure 18.12 shows five ships sailing between Beaumont and Punta Arenas. Planes and railcars of the same type have the same operating characteristics; multiple assets of the same type moving between the same origin and destination represent one route. Rows six and seven of Figure 18.12 show a total of five trains hauling cargo from Fort Hood, TX, to the seaport at Beaumont; the different listings indicate different types of rail cars used for each group of trains.

A key measure of effectiveness for the deployment schedule is the utilization of the transportation assets. After the scheduling procedure is complete, DSAT calculates the utilization of all assets on each route. First, the total amount of weight moved over a specific route is computed and displayed as depicted in Figure 18.12. Next, based on the operating characteristics of each asset type operating on a route, the possible number of round trips completed by a single asset is computed. The asset-route weight utilization percentage reflects the average percentage of asset weight capacity utilized in each trip, assuming that all assets assigned to the route completed the maximum possible number of round trips. The volume utilization percentage is computed similarly. The "volume" utilization for ships is actually the utilization of the ship's loading area since the capacity of ships used for moving military equipment is normally expressed in terms of surface loading area. The weight volume (area for ships) utilization of assets on each route is listed in the bottom table of Figure 18.12. Graphs are generated for the user to view the utilization of assets on a route over time. An example is provided in Figure 18.14. This graph represents the utilization of the surface loading area of one light-medium ship, roll-on, roll-off (LMSR). The surface area, adjusted for necessary tie-down equipment, was effectively 100% utilized by the equipment loaded on this ship from the scheduling procedure.

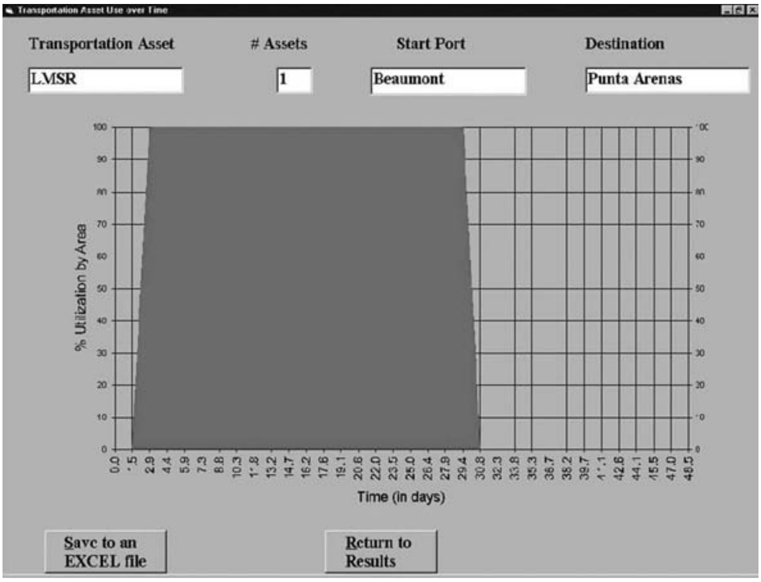


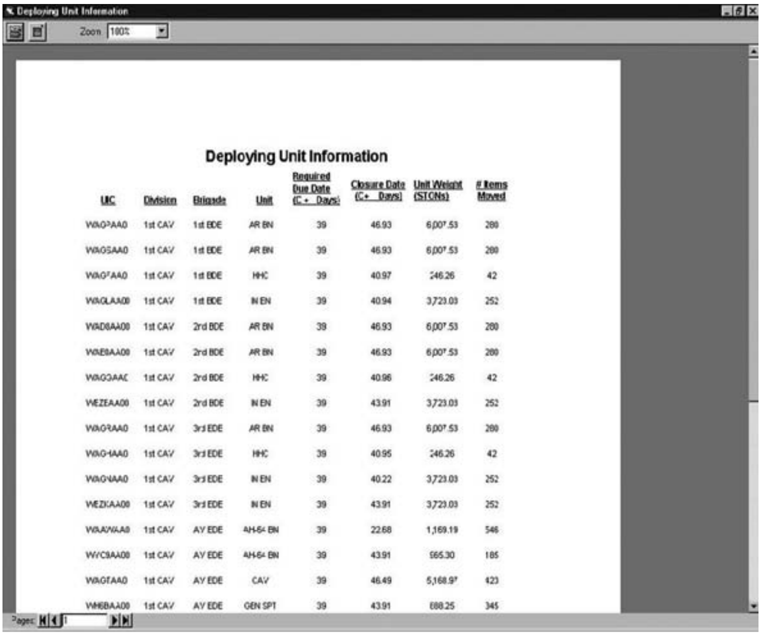
FIGURE 18.14 Example graph of ship loading area utilization.

18.4.2 Data Reports of Deployment Information

Deployment planners need information on the schedule of unit equipment movement and transportation asset use in an understandable format. DSAT generates several reports that planners can view, print and/or save. The data displayed in these reports is generated through database queries and by using the output information from the scheduling routine. A sample of the Deploying Unit Information Report is provided in Figure 18.15.

A brief description of each report is provided below:

- **Deployment schedule by unit** details how a unit moved from its start base to the POD(s). If any items of equipment were on a particular transportation asset, the unit's UIC is listed along with the C-date/time the asset started loading and when the asset completed unloading at its destination. The asset name and number are also provided.
- **Deployment schedule by unit and item of equipment** details how each item of equipment in each unit moved from its start base to the POD(s). The C-date/time when loading begins, the origin, destination, C-date/time when unloading is complete and the transportation asset name and number are provided for each movement an item of equipment makes.
- **Transportation asset schedule** chronologically lists the C-date/time of each load and unload activity for each transportation asset used in the deployment.
- **Transportation asset input and use information** lists the key operating characteristics for each type of transportation asset used in the deployment as well as the total weight carried by each asset type and its percent utilization by both volume and weight.
- **Deploying unit information** details the required delivery date (RDD), actual closure date, total weight of unit equipment (in STONs) and number of equipment items for each unit.
- **Port information** lists the port facility input characteristics, e.g. the MOG and number of runways for an airport and the amount of cargo processed through the port (STONs).
- **Items needing special rail clearance** lists any items of equipment detected as exceeding the dimensions of railcars chosen during the feasibility checks.
- **Equipment removed from deployment** If the feasibility checks identify equipment not fitting on any transportation asset selected, the user can elect to remove those items from the deploying equipment list. Details on these items are in this report.



Deploying Unit Information

UIC	Division	Brigade	Unit	Required Due Date (C+ Days)	Closure Date (C+ Days)	Unit Weight (STONs)	# Items Moved
WAG?AA0	1st CAV	1st BDE	AR BN	39	46.93	6,007.53	280
WAG5AA0	1st CAV	1st BDE	AR BN	39	46.93	6,007.53	280
WAG7AA0	1st CAV	1st BDE	HHC	39	40.97	246.26	42
WAGLAA00	1st CAV	1st BDE	IN BN	39	40.94	3,723.03	252
WAD8AA00	1st CAV	2rd BDE	AR BN	39	46.93	6,007.53	280
WAE8AA00	1st CAV	2rd BDE	AR BN	39	46.93	6,007.53	280
WAG0AAC	1st CAV	2rd BDE	HHC	39	40.96	246.26	42
WEZEAA00	1st CAV	2rd BDE	IN BN	39	43.91	3,723.03	252
WAGRAA0	1st CAV	3rd BDE	AR BN	39	46.93	6,007.53	280
WAG4AA0	1st CAV	3rd BDE	HHC	39	40.95	246.26	42
WAGNAA0	1st CAV	3rd BDE	IN BN	39	40.22	3,723.03	252
WEZKAA00	1st CAV	3rd BDE	IN BN	39	43.91	3,723.03	252
WAAWAA0	1st CAV	AV BDE	AH-64 BN	39	22.68	1,169.19	546
WYC9AA00	1st CAV	AV BDE	AH-64 BN	39	43.91	565.30	185
WAGTAA0	1st CAV	AV BDE	CAV	39	46.49	5,168.97	423
WH6BAA00	1st CAV	AV BDE	GEN SPT	39	43.91	688.25	345

FIGURE 18.15 Sample deploying unit information report.

18.5 Conclusions

DSAT is a personal computer based deployment planning tool that provides maximum flexibility to deployment planners. Deployment planners can model various deployment scenarios and rapidly create readable "optimal" deployment schedules and conduct valuable sensitivity analysis of various aspects of the deployment problem. DSAT occupies an important niche in the realm of existing deployment planning tools. It quickly provides the detailed information needed to create a deployment schedule. DSAT also provides the processing speed necessary for sensitivity analysis not provided by more detailed scheduling models.

References

1. Chamberlain, R. G. 1996. Calculating Distances on the Surface of the Earth [Online]. Available: http://www.census.gov/cgi-bin/geo/gisfaq?O5.1.
2. Department of Defense Almanac. 2001. Unified Combatant Commands [Online]. Available: http://www.defenselink.mil/pubs/almanac/unified.html.
3. Greer, A., and D. Hancox. 1991. *Tables, Data and Formulae for Engineers and Mathematicians*. Stanley Thornes Publishers Ltd, ISBN 0-7487-0077-3.
4. Hodgson, T. J., D. Cormier, A. J. Weintraub, and A. Zozon. 1998. Satisfying due dates in large job shops. *Management Science*, 44(10), 1442–1446.
5. Illingworth, J. 1999. A simulation based modeling of military transportation applications: A combat logistics and scheduling system for deployment of United States Armed Forces worldwide. Unpublished Masters' Thesis, North Carolina State University, Raleigh, NC.
6. Hodgson, T. J., O. Spivey, T. E. Trainor, and M. K. Williams. 2001. The Deployment Analysis Network Tool Enhanced (DANTE). *Military Operations Research*, 6(3).
7. Melendez, B. S. 2002. On scheduling delivery in a military deployment scenario. Unpublished Ph.D. Dissertation, North Carolina State University, Raleigh, NC.

8. MTMCTEA. 2000. Modeling and Simulation [Online]. Available: http:\\www.tea.army.mil\dpe\modeling.html.
9. North, C. 2001. Lat/Lon to Distance! [Online]. Available: http://home.eol.ca/~cnorth/frame_11.html.
10. Sinnott, R. W. 1984. Virtues of the Haversine. *Sky and Telescope*, 68(2).
11. Thoney, K.A. 2000. Scheduling batch processing machines in job shop scenarios. Unpublished Ph.D. Dissertation, North Carolina State University, Raleigh, NC.
12. Thoney, K. A., T. J. Hodgson, R. E. King, M. Taner, A. Wilson. 2002. Satisfying due-dates in large multi-factory supply chains. *IIE Transactions,* 34(9).
13. Trainor, T. E. 2001. Scheduling military deployment. Unpublished Ph.D. Dissertation, North Carolina State University, Raleigh, NC.
14. Williams, E. 2001. Aviation Formulary Vl.30 [Online]. Available: http://www.best.com/~williams/avform.htm#GCF.

19

The Deployment Analysis Network Tool Extended (DANTE)

Thom J. Hodgson
North Carolina State University

W. Owen Spivey
Combat Development Engineering Field Office

Tim E. Trainor
United States Military Academy

Michael K. Williams
Transportation Engineering Agency

19.1 Introduction

The Deployment Analysis Network Tool Extended (DANTE) is an analysis tool used by the Transportation Engineering Agency of the Military Traffic Management Command to examine the logistics of large-scale troop deployments. DANTE models the flow of troops and equipment from air and seaports of embarkation (APOE, SPOE), through air and sea corridors to air and seaports of debarkation (APOD, SPOD), then through forward transportation corridors (air, road, rail, or waterway) to the final assembly area. DANTE originally was develop as a software tool to satisfy the analysis needs of an Army Science Board (ASB) study involving strategic mobility concepts in 2025 time frame ASB 1999). There was a need for determining the sensitivity of the *time to close the force* to various alternative means of transportation and to various alternative logistic capacities and infrastructures. Due to the requirement for analyzing a very large number of scenario alternatives, the tool had to be capable of modifying scenarios quickly and then finding optimal deployment plans in near real time. DANTE uses as its basis a network depiction of the deployment process developed for the Army After Next (AAN) war game of 1999. The process is modeled as a time-phased network flow. With an objective of minimizing the *time to close the force,* the model is optimized using a min-cost network flow algorithm (Fulkerson 1961; Minty 1960).

DANTE does not provide detailed Time Phased Force and Deployment Data (TPFDD). Rather it flows troops and equipment through the deployment network with appropriate time delays for travel, loading/unloading, processing through logistic nodes, and queuing of troops and equipment at choke points when necessary.

The Military Traffic Management Command–Transportation Engineering Agency (MTMC-TEA) developed the notional scenario used in the original ASB study. The office of the Army Deputy Chief of Staff for Operations (ODCSOPS) developed a multidivision joint force structure that was used in the

ASB study to evaluate the concept of strategic maneuver. DANTE was validated against the Joint Flow and Analysis System for Transportation (JFAST) (Oak Ridge National Laboratories 2000) by MTMC-TEA using a number of varying scenarios and was found to give results that were consistent with that model. JFAST is a detailed, simulation based planning model used to estimate deployment transportation requirements and unit closure time. However, setting up a scenario for JFAST is time consuming. It can take an hour or more to run. Consequently, it does not lend itself well to the requirements of sensitivity analysis where large numbers of scenarios to be evaluated.

In the following sections, the general framework and network flow structure of DANTE is developed. Then, a select number of sample runs are presented to illustrate some of the analysis capabilities of the tool. Finally, we conclude the chapter with some observations and recommendations as to how DANTE can be used to extend the analytic capability beyond that presented here. The DANTE program is available from MTMC-TEA.

19.2 The Model

The fundamental network structure of DANTE is depicted in Figure 19.1. However, the actual network model is somewhat more complicated. DANTE is essentially a pipeline model in which equipment and troops are "flowed" through the network to their final destination (the staging area). Various nodes of the model (airports and seaports) are limited in their capacity to process people and equipment. These capacity limitations are the actual (or forecast) capacities of the facilities to be used, and are expressed in tons/day. Various transportation links (air routes, sea routes, roadways, and railways) are also limited in their capacities. These capacity limitations are a function of the specific assets available (number of and type of aircraft, number of and type of ships), the distance involved in the deployment, and/or the

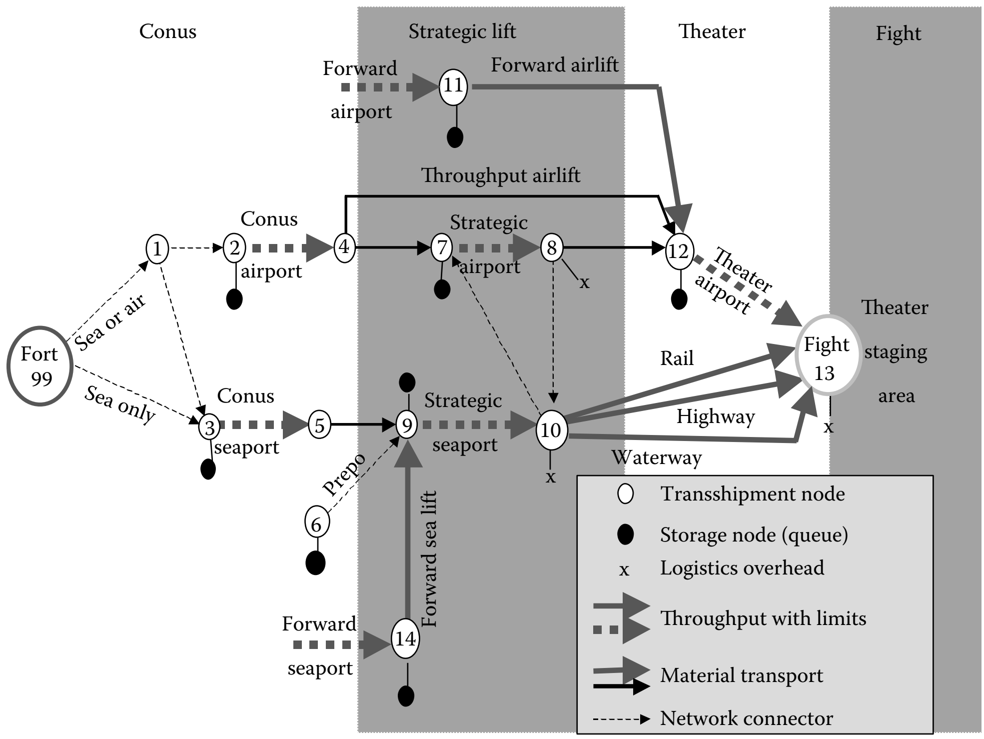


FIGURE 19.1 DANTE aggregate schematic.

capacity of the roadways and railways involved. As a result, material may be delayed (queued) at certain points in the network. In addition, the flow of material within the transportation links of the model is delayed by the length of time associated with the particular link (i.e. flight, sea, rail, and road times).

The network illustrated in Figure 19.1 represents a single snapshot of the DANTE network flow, which evolves in discrete six-hour segments. Thus, one could conceptualize the entire model as a multilevel stacked sequence of identical snapshots, each of which represents the network at a particular time. Thus, a ship leaving the SPOE at time equal 0.0 (node 5, level 0 in the stacked network), with a four-day travel time arrives at the SPOD at time equal 4.0 (node 9, level 16 in the stacked network). Each possible passage is represented by an arc connecting the nodes along with an associated spanning time. Figure 19.2 shows a partial side view of the stacked network with the time dimension exposed. Flow moves left to right from the APOE, to the APOD, and on to the staging area, and flow moves upward in time depending on the duration of the process.

Since specific items of equipment flow through the network as tons, they are blurred in the process, and flow relative to sea-borne and air-borne resources can be optimized. As a side benefit, the time consuming process of developing the TPFDD is not necessary. For a specific set of inputs (i.e. required troops and equipment to be moved, and the capacities of the available transportation equipment and logistics infrastructure), DANTE identifies an allocation of troops and equipment to the various available transportation assets that minimizes the time to close on the staging area. Optimization is performed using a min-cost network flow algorithm.

DANTE does not directly represent several significant planning factors such as transportation, holding and handling costs, threats to transportation routes, and command and control factors. It can, however, be used to model the effects of these factors by limiting the capacities of specific links in the network. The issue, then, is the sensitivity of the model to deterioration in the capacity of the various elements (APOE's, SPOE's, APOD's, SPOD's aircraft, road nets, etc.).

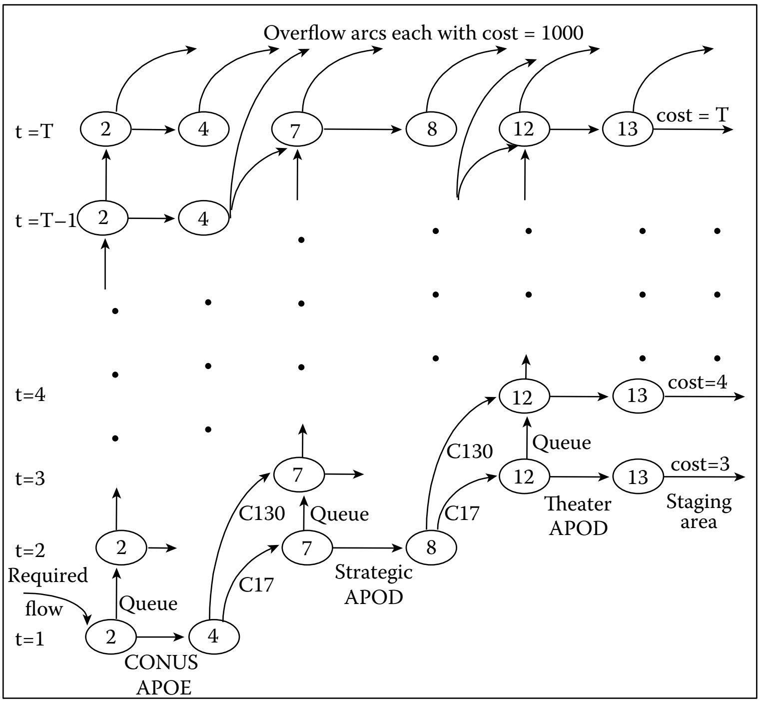


FIGURE 19.2 Partial side view of DANTE network.

Since troops and equipment flow to their destinations represented as aggregate tons moved, they lose their identities in DANTE. Furthermore, individual ports are aggregated during each phase of a deployment scenario. For example, in Figure 19.1 nodes 2 and 4 represent the aggregate capacity of all airports used in the continental United States (CONUS) as APOE's to deploy forces. If multiple ports are use in other phases of the deployment, their capacities are likewise aggregated. This introduces a bias into the model results in that, for a given scenario, the time to close solution may be less than that achievable in reality. However, the network flow optimizer allocates personnel and equipment in such a way as to utilize most efficiently the transportation and port resources available. If the model indicates that a particular scenario cannot be executed successfully within a required time frame, that scenario would not be possible in the real world either. Thus, although the optimal time to close solutions have the potential to be overly optimistic, the fidelity of the model to feasibility is robust.

In order to use the model effectively, specific scenarios must be carefully constructed and evaluated to ascertain the sensitivity of the time to close on the staging area to the available transportation asset levels and infrastructure capabilities. This particular sensitivity captures the tradeoffs between methods of delivery to the theater for a given level of assets needed.

As mentioned above, the model is capacitated at a number of points. The CONUS APOE (arc 2–4), the CONUS SPOE (arc 3–5), the strategic APOD (arc 7–8), the strategic SPOD (arc 9–10), the forward APOE (arc 11–12), the forward SPOE (arc 14–9) and the theater airport (arc 12–13) all are limited to a given tons per day in their capacity to process personnel and equipment. The transportation arcs in the network: strategic airlift (arcs 4–7 and 4–12), strategic sealift (arc 5–9), theater airlift (arc 8–12), forward airlift (arc 11–12), forward sealift (arc 14–9), highway (arc 10–13), rail (arc 10–13) and waterway (arc 10–13) are limited in their capacity (tons per day) to transport personnel and equipment based on the transportation assets assigned to the arch. The capacity of an arc is proportional to the number of assets assigned and the unit asset capacity. It is inversely proportional to the sum of the round trip travel time, material loading time, and unloading time (see the formulas in the "Input parameter" section). In addition, flow is delayed along these arcs to represent transportation times. The delay on an arc is proportional to the travel distance and inversely proportional to the asset speed, plus the material loading and unloading times (see the formulas in the "Input parameter" section).

19.3 Model Structure

As noted above, the model is structured as a time phased network with a basic time increment of six hours. For instance, equipment leaving the SPOE (node 5) in the middle of day 3 (i.e. day = 3.5) for a trip of 5.25 days, arrives at the strategic SPOD (node 9) in the last quarter of day 8 (i.e. day = 8.75). In order to achieve the objective of minimizing the time to close the force, costs are assigned to certain arcs as shown in Figure 19.2. Using these costs, linear programming provides the allocation of flow in the model. Costs are assigned as follows:

(1) A cost of 1000 is assigned to each unit of flow on overflow arcs (i.e. arcs that go beyond the maximum time allowed by the model). These arcs are use only if the delivery schedules cannot be met through normal flow. This is simply a practical facet that insures the "mathematical feasibility" of the model.
(2) A linearly increasing cost is assigned to the flows out of node 13. The cost increases each period. Since flow in earlier periods is cheaper, this insures that the time to closure of the force is minimized.

The most important result of running the model is a measure of how quickly the scenario can be brought to closure. For further analysis, indications of bottlenecks are given with graphs showing where backlogging has occurred (see Figures 19.4 and 19.5). Critical flows in each time period are output for more detailed analysis (see Figure 19.4).

19.4 Input Parameters

An example of the form for entering model input parameters can be seen in Figure 19.3. The following defines the parameters used (MTMC-TEA 1997a and b).

Sea distance (SPOE): The distance (nautical miles, nm), from the SPOE to the SPOD.

Forward sea distance: The distance (nm), from the forward SPOE to the SPOD.

Number of ship types: Number of different ship types in the data.

#Ships type *i* (strategic): The number of ships of type *i* available to sail from CONUS.

#Ships type *i* (forward): The number of ships of type *i* available to sail from the forward base.

Ship speed *i*: The speed (knots), that ships of type *i* travel.

Ship capacity *i*: The carrying capacity (tons), of a ship of type *i*

Ship load time *i*: Type *i* ship load time (hours).

Ship unload time *i*: Type *i* ship unload time (hours).

Sea capacity *i*: Maximum number of tons/hour carried in strategic sealift by ships of type *i* (calculated, variables not subscripted for simplicity).

(Sea cap = (# ships) * (ship cap)/((2 * sea distance)/(ship speed) + (ship load time) + (ship unload time))

Sea days *i* from CONUS: The time to transport equipment from the CONUS seaport to the strategic seaport by ships of type *i* (calculated, no subscripts used).

Sea days = (sea distance)/(ship speed) + (ship load time) + (ship unload time).

Strategic air distance: The distance (nm), from CONUS APOE to the strategic APOD.

Through air distance: The distance (nm), from CONUS APOE to the theater APOD.

Forward air distance: The distance (nm), from forward APOE to the theater APOD.


Form1

Enter Applicable Shipping Input Data

SEA DISTANCE SPOE->SPOD (NM)	9010
FORWARD SEA DISTANCE (NM)	5121
NUMBER-OF-SHIP-TYPES——(#)	9
LMSR (STRATEGIC)(#)	10
LMSR (FORWARD) (#)	0
LMSR SPEED (KNOTS)	24
LMSR CAPACITY (TONS)	17177
LMSR LOAD TIME (HOURS)	48
LMSR UNLOAD TIME (HOURS)	36
FSS (STRATEGIC)(#)	8
FSS (FORWARD) (#)	2
FSS SPEED (KNOTS)	28
FSS CAPACITY (TONS)	10630

NOTE: ALL INPUT VALUES MUST BE INTEGERS. When changing data, you must use the mouse to click off the last row changed in order for the changes to take affect. Simply click on the cell of another row after your last change.

Click to go to Air Inputs

FIGURE 19.3 Sample input parameters form.

Theater air distance: The distance (nm), from strategic APOD to the theater APOD.

Number of aircraft types: Number of different aircraft types in the data.

#Aircraft type *j* (strategic): The number of aircraft of *j* available to be used for transport from the CONUS APOE to the strategic APOD.

#Aircraft type *j* (through): The number of aircraft to type *j* available to be used for transport from the CONUS APOE to the theater APOD.

#Aircraft type *j* (forward): The number of aircraft of type *j* available to be used for transport from the forward APOE to the theater APOD.

#Aircraft type *j* (theater): The number of aircraft of type *j* available to be used for transport from the strategic APOD to the theater APOD.

Aircraft speed *j*: The speed (knots), that aircraft of type *j* travel.

Aircraft capacity *j*: The carrying capacity, in tons, of aircraft of type *j*.

Aircraft load time *j*: Type *j* aircraft load time (hours).

Aircraft unload time *j*: Type *j* aircraft unload time (hours).

Strategic-air capacity *j*: Maximum number of tons/hour carried in strategic airlift from CONUS APOE to the strategic APOD by aircraft type *j* (calculated, variables not subscripted for simplicity).

Strat air cap = (# a/c) * (a/c cap)/((2 * strat air distance)/(a/c speed) + (a/c load time) + (a/c unload time)).

Strategic-air time *j*: The time (hours) to transport equipment from the CONUS APOE to the strategic APOD by aircraft type *j* (calculated, no subscripts used).

Strat air time = (strat air distance)/(a/cspeed) + (a/c load time) + (a/c unload time)

Through air capacity *j*: The maximum number of tons/hour that can be carried in military airlift from the CONUS airport to the theater airport by aircraft type *j* (calculated, no subscripts used).

Thru air cap = (# a/c) * (a/c cap)/((2 * thru air distance)/(a/c speed) + (a/c load time) + (a/c unload time)).

Through-air time *j*: The time to transport equipment from the CONUS APOE to the theater airport by aircraft type *j* (calculated, no subscripts used).

Thru air time = (thru air distance)/(a/c speed) + (a/c load time) + (a/c unload time).

Theater air capacity *j*: The maximum number of tons/hour that can be carried in military airlift from the strategic APOD to the theater airport by aircraft type *j* (calculated, no subscripts use).

Theater air cap = (# a/c) * (a/c cap)/((2 * theater air distance)/(a/c speed) + (a/c load time) + (a/c unload time)).

Theater-air time *j*: The time to transport equipment from the strategic APOD to the theater airport by aircraft type *j* (calculated, no subscripts used).

Theater air cap = (theater air distance)/(a/c speed) + (a/c load time) + (a/c unload time).

CONUS seaport capacity: The maximum number of tons/day that can be processed through the CONUS SPOE and loaded on ships.

CONUS airport capacity:	The maximum number of tons/day that can be processed through the CONUS APOE and loaded on ships.
Forward seaport capacity:	The maximum number of tons/day that can be processed through the forward SPOE and loaded on ships.
Forward airport capacity:	The maximum number of tons/day that can be processed through the forward APOE and loaded on aircraft.
Strategic airport capacity:	The maximum number of tons/day that can be offloaded from strategic aircraft and processed through the strategic APOD.
Strategic seaport capacity:	The maximum number of tons/day that can be offloaded from strategic sealift and processed through the strategic SPOD.
Theater road time:	The transport time in days by road from the strategic APOD (or SPOD) to the staging area.
Theater road capacity:	The maximum number of tons/day that can be carried by road from the strategic APOD (or SPOD) to the staging area.
Theater rail time:	The transport time in days by rail from the strategic APOD (or SPOD) to the staging area.
Theater road capacity:	The maximum number of tons/day that can be carried by rail from the strategic APOD (or SPOD) to the staging area.
Theater water time:	The transport time in days by waterway from the strategic APOD (or SPOD) to the staging area.
Theater water capacity:	The maximum number of tons/day that can be carried by waterway from the strategic APOD (or SPOD) to the staging area.
Theater airport capacity:	The maximum number of tons/day that can be offloaded from military aircraft and processed through the theater APOD.

19.5 Lift Profile

The lift profile is a way of detailing the availability of movement requirements. Specifically, DANTE takes as input the delivery schedules of equipment to the CONUS SPOE and APOE and to the forward SPOE and APOE. This is referred to as the available to load date (ALD). The user specifies the total number of tons that are required to go by sea at each SPOE, along with the delivery schedule of pre-positioned (PREPO) equipment to the strategic SPOD. Finally, the user can specify the logistics overhead usage in short tons (STONs) at the strategic SPOD and APOD, and at the theater APOD. The logistics overhead is simply the amount of equipment that will remain at or be consumed by these locations as part of their port support activity.

19.6 Using DANTE

The deployment parameters are input through a graphic user interface (GUI) which is programmed in VISUAL BASIC©. It allows the user to change inputs by simply typing over the previous values. The input data is maintained in an ACCESS© database and is formatted for solving by the network optimizer through structured query language (SQL). A sample screen for editing input parameters is shown in Figure 19.3.

The column headers of Figure 19.4 provide the node numbers for reference to the network diagram. An asterisk next to the value in a field indicates the arc is operating at capacity during that time period. This is a key element in performing sensitivity analysis (see example in the next section). The user can view graphically the backlog over time at any of the bottleneck nodes as illustrated in Figure 19.5.

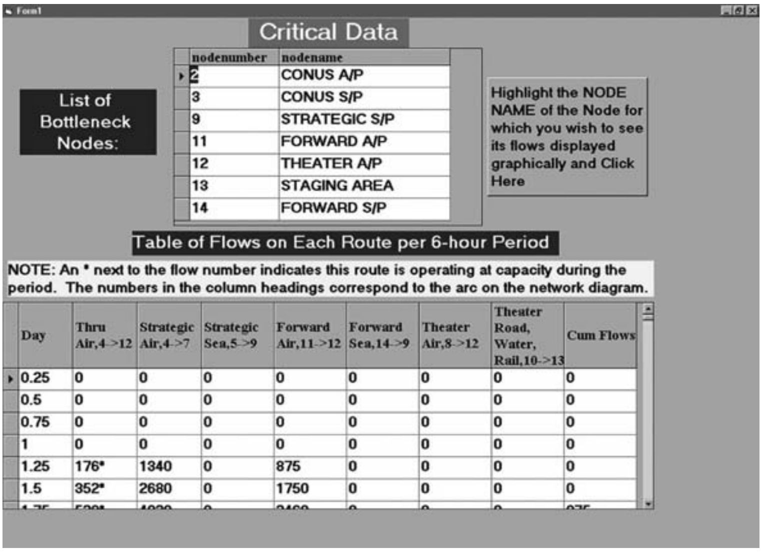


FIGURE 19.4 DANTE critical data output display.

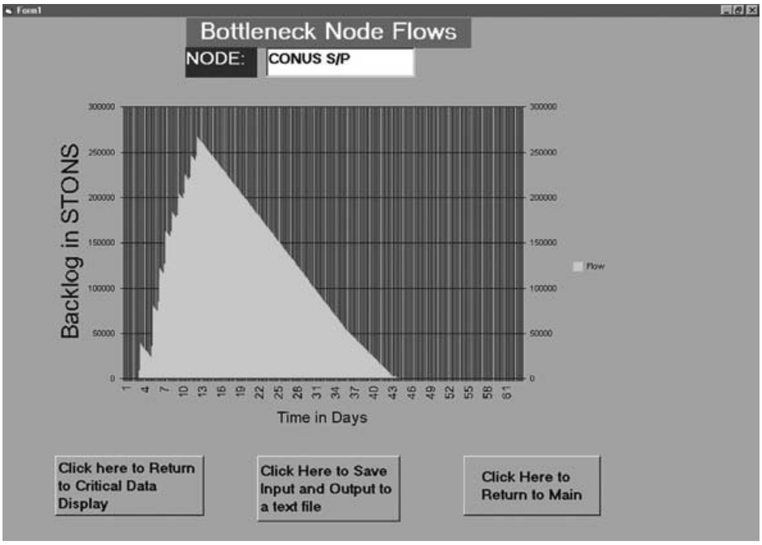


FIGURE 19.5 Graph of equipment backlog at CONUS SPOE node.

The vertical axis represents equipment, in STONs, while the horizontal axis is days. The graph shows the CONUS SPOE node backlog increasing to its maximum level at day 12, which is the last day cargo arrives at this port for movement. The backlog at the aggregate CONUS SPOE node is finally reduced to zero after approximately 44 days. The ease of setting and changing deployment parameters and the speed of the network optimizer (less than 20 seconds solution time on a 150 megahertz PC) make DANTE an excellent tool for performing sensitivity analysis. Note that DANTE is a high level model that provides aggregate type detail but does not deal with individual items—all cargo is treated as tons. The sensitivity analysis is demonstrated through the following example deployment scenario.

19.7 Sensitivity

The goal of this example analysis is to show how the DANTE output can be used to determine the main factor(s) slowing the overall deployment time. The sample input and output (Figures 19.3 through 19.5) is from an excursion to Omnia, a notional country in western Africa. The parameter set describing this excursion is contained in the appendix at the end of this chapter. A battle force that is air and/or sea transportable is being sent from CONUS, Europe and PREPO stocks with the delivery schedule to the ports specified in the lift profile. The characteristics of the air and sea transportation resources available along with the port capacities and theater time constraints are also defined in this chapters appendix.

The following discussion puts the parameters into context. This deployment requires moving the following aggregate amount of equipment and supplies:

- 350,606 STONs by ship from CONUS
- 85,929 STONs by either ship or aircraft from CONUS
- 44,549 STONs by ship from forward based forces in Europe
- 27,924 STONs by either ship or air from forward based forces
- 72,667 STONs of PREPO in theater

This roughly equates to a deployment of two heavy Army divisions plus an airborne or air assault division from CONUS. The CONUS deploying forces also include Special Operations and Marine units, and support units and equipment for Air Force combat aircraft and Army logistical support units. This scenario also includes deploying roughly one Army heavy brigade from forward based units in Europe along with logistical support and Air Force support units. The PREPO force includes an Army heavy brigade, support forces and cargo packages for logistical support.

Moving such a large amount of cargo quickly requires a great deal of logistical infrastructure. The ports used in this scenario include both military and civilian facilities. It is assumed that the military can use approximately 50% of the available infrastructure at the civilian ports included. The CONUS airport throughput (20,000 STONs/day) represents a combined capacity of five airfields similar in size to the following:

- Dallas-Fort Worth International, Texas
- Robert Gray Army Airfield (AAF), Fort Hood, Texas
- Hunter AAF, Fort Stewart, Georgia
- Louisville International, Kentucky
- Pope Air Force Base, North Carolina

The CONUS seaport throughput capacity (165,000 STONs/day) represent the combined capacity available to the military at the following ports (National Imagery and Mapping Agency 1998):

- Beaumont, Texas
- Savannah, Georgia
- Newport News, Virginia
- Bayonne, New Jersey

The forward seaport throughput (60,000 STONs/day) represents using the capacity available to the military at several ports with good facilities in major industrial areas of Europe, for example Rotterdam. Similar to the seaport use, the forward airport throughput (14,000 STONs/day) represents the combined throughput available to the military associated with several commercial airfreight terminals in industrial areas of Europe. Ramstein Air Force Base, Germany is also included in the forward airport capacity.

Since the deployment is to a notional country in Africa, the available infrastructure capacities reflect estimates of the facilities available in an emerging nation (Central Intelligence Agency 1999). The limited strategic seaport capability (18,000 STONs/day) reflects using less than 100% of the estimated capacity of two major ports in an emerging nation. The limited infrastructure available in emerging nations is also modeled in the strategic airport throughput (5361 STONs/day). This is the combined throughput for two moderately sized international airfields capable of handling C-5 and B-747 aircraft. Airfields being used by the Air Force for fighter, tanker, and other support are not included in this estimate. The theater airport capacity (3500 STONs/day) reflects the availability of several small forward landing strips near the staging area capable of handling C-130 and C-17 cargo aircraft. Omnia is assumed to have a limited highway, rail and waterway network. The capacity estimates used initially are 8000 STONs/day for theater road and 5000 STONs/day for theater rail. Furthermore, the deployment plan does not include moving cargo in theater using inland waterways. The infrastructure constraints in emerging nations, as demonstrated in this example, can impact the speed with which military forces deploy.

For this scenario, DANTE produces an optimal solution of 63.0 days to close all equipment on the staging area. Backlogs occur at the CONUS air and seaports, the forward air and seaports, the theater airport and the strategic seaport. Figure 19.5 shows a significant backlog at the CONUS SPOE. Viewing the graphs for the other bottleneck nodes quickly shows the CONUS SPOE backlog is the greatest over time. Table 19.1 provides the day at which the backlog at each bottleneck node peaked and was cleared, as shown by the bottleneck graphs.

From Table 19.1 it is clear that the focus should be on the shipping aspect of the deployment scenario to improve closure time, since the CONUS and Forward SPOE backlogs are cleared later than the other ports. From the flow on the transportation arcs, it is easy to see that the shipping backlog is due to lack of transportation assets (i.e. transportation arc flows are at capacity). Specifically, the input parameters specify the CONUS SPOE has a throughput capacity of 165,000 STONs/day while the Forward SPOE capacity is 60,000 STONs/day. Since only 350,606 STONs need to be moved from the CONUS SPOE, theoretically this node could be cleared in three days with sufficient shipping capacity. Likewise, only 44,549 STONs need to be moved from the Forward SPOE, indicating this node could be cleared in less than one day if the requisite shipping capacity existed. Furthermore, if one were to scroll down the output table of flows on routes (Figure 19.4), the strategic sea route (arc 5–9) operates at capacity from day 3.25 through day 35.75. The forward sea route (arc 14–9) also operates at capacity from day 3.25 to day 17.0. The combination of the backlog at these nodes, along with the capacitated arcs leaving these nodes, indicates the backlog could be reduced, and hence deployment speed improved, by adding shipping capacity to these routes.

The adjustments that we now make to the deployment parameters are simply to highlight the insights that can be gained using DANTE. Obviously, determining the feasible set of deployment parameters is not a simple task. It takes coordination and planning by key military decisionmakers. The goal here is simply to demonstrate how DANTE can quickly and accurately quantify the deployment planner's insights in improving overall deployment closure time.

Increasing the number of container ships delivering equipment from the CONUS and Forward SPOE's to the Strategic SPOD is a good option to reduce the backlog, assuming that the equipment can be placed in containers. Improving the cargo handling operations at seaports can also reduce backlogs. We first increase the number of container ships departing the CONUS SPOE from five to 40 and allocate an additional 20 container ships to the route departing the Forward SPOE (Military Sealift Command 2000). Furthermore, we assume that additional resources (e.g. material handling equipment) can be allocated to seaports allowing container ship load and unload times to be reduced by on-third, while

TABLE 19.1 Day of Backlog Peak and Clearance at Ports

Node	Day Peaked	Day Cleared
CONUS APOE	5	16
CONUS SPOE	12	44
Strategic SPOD	4	13
Forward APOE	4	12
Forward SPOE	5	21
Theater APOD	5	6

TABLE 19.2 Day of Backlog Peak and Clearance at Ports (After the 1st Adjustment)

Node	Day Peaked	Day Cleared
CONUS APOE	5	16
CONUS SPOE	12	21
Strategic SPOD	4	12
Forward APOE	4	11
Forward SPOE	4	10
Theater APOD	5	6

all other ship load and unload times are reduced by one-quarter. Implementing these changes within DANTE yields an improved closure time of 47.25 days (a reduction of 25%). Table 19.2 provides the day backlogs peaked and were cleared at each port after implementing the changes in the deployment scenario.

Improving cargo handling speed and increasing the shipping capacity leaving the CONUS and Forward SPOEs decreased the backlogs to 21 days and 10 days, respectively, while not adversely affecting the backlogs at downstream nodes.

The dramatic increase in the shipping capacity intuitively should have a greater impact on the closure time. Node backlogs disappear by day 21; however the force does not close until day 47.25. DANTE output indicates that infrastructure constraints now show the deployment. Specifically, scrolling through the table of flows on each route (Figure 19.4), the theater road, rail and waterway routes (arcs 10–13) operate at capacity for many periods from day 3.25 through day 45.25. No other routes have significant capacity problems. Hence, the limited infrastructure in country is clearly constraining this deployment scenario.

We next adjust to the deployment parameters to focus on increasing the capacity to move cargo in theater. First, we increase the theater road capacity to 20,000 STONs/day by negotiating for additional highway use and contracted line haul and by moving military transportation units up in the deployment arrival timeline. By selecting better ground shipping routes the road time to the staging area is also reduced from one day to 0.75 days. Contracting for additional railcars increases the theater railway capacity to 12,500 STONs/day. We then modify the deployment plan to incorporate movement of cargo over inland waterways with a daily capacity of 5000 STONs. Finally, we employ more forward landing strips near the TAA capable of handling C-17 and C-130 aircraft, thereby increasing the theater APOD capacity to 8000 STONs/day. Implementing these adjustments in the deployment parameters, and re-running DANTE, yields a closure time of 45.25 days (another 4.2% decrease). This iterative process could be continued for further reductions in time to close.

While DANTE lacks the overall fidelity of a detailed planning tool like JFAST, DANTE can help a planner quickly identify and quantify various changes needed in deployment parameters to improve closure time. For the example used here, the results of the original DANTE run indicated that the shipping capacity leaving the CONUS and Forward SPOE's was the main factor slowing the overall deployment time. The impact on closure time directly associated with adding container ships and improving cargo-handling procedures was easily quantified. These improvements then further highlighted an infrastructure problem in theater. DANTE was then used to quantify the impact of certain improvement of theater infrastructure supporting the deployment.

19.8 Conclusion

DANTE also could be used to compare alternatives for increasing shipping capacity or improving theater infrastructure for this example. The speed at which the deployment parameters can be changed and the model resolved make DANTE an effective tool for conducting sensitivity analysis of deployment scenarios. MTMC-TEA routinely uses DANTE to assess impacts of infrastructure changes, force modernization initiatives, and new lift platforms on force closure time. DANTE results are then used to determine which alternatives require further analysis with more complex models and simulations.

The general contribution of this effort has been to facilitate the sensitivity analysis of deployment scenarios for the purposes of providing quick turnaround answers to support Army force modernization initiatives and the evaluation of new strategic lift platforms. It has been used in support of two ASB studies. Current development of DANTE includes detailed modeling of multiple APOE, VSPOE, APOD, SPOD scenarios and the optimal allocation of transportation resources on the various routes. In addition, a deployment analysis system that bridges the gap between DANTE and JFAST is under development.

References

Army Science Board. 1999. *Strategic Maneuver: Report to Chief of Staff Army*. National Academy for Science and Engineering. Beckman Center, University of California, Irvine.

Central Intelligence Agency. 1999. *The World Factbook* [Online]. Available: http://www.odci.gov/cia/publications/factbook/.

Fulkerson, D.R. 1961. An out-of-kilter method for minimal cost flow problems. *Journal of the Society for Industrial and Applied Mathematics*, 9(1), 18–27.

Military Sealift Command of the United States Transportation Command. 2000. *Surge Project Office of the Sealift Program* [Online]. Available: http://www.msc.navy.mil/PM5/surge.htm.

Military Traffic Management Command, Transportation Engineering Agency. 1997a. *Deployment Planning Guide: Transportation Assets Required for Deployment*. MTMCTEA Reference 97-700-5, 2.

Military Traffic Management Command, Transportation Engineering Agency. 1997b. *Logistics Handbook for Strategic Mobility Planning*. MTMCTEA Reference 97-700-2.

Minty, G.J. 1960. Monotone networks. Proceedings of the Royal Society of London, Ser. A., 257, 194–212.

National Imagery and Mapping Agency. 1998. *World Port Index, Publication 150* [Online]. Available: http://164.214.12.145/pubs/pubsjwpisections.html#viewpub.

Oak Ridge National Laboratories. 2000. *Joint Flow and Analysis System for Transportation (JFAST)* [Online]. Available: http://www.JFAST.org.

Appendix

Initial Deployment Parameters for the Notional Omnia Scenario

	Value
Sea Transportation Parameters	
Sea distance SPOE to SPOD (nm)	9010
Forward sea distance (nm)	5121
Number of ship types (#)	6
LMSR (strategic) (#)	10
LMSR (forward) (#)	0
LMSR speed (knots)	24
LMSR capacity (tons)	17177
LMSR load time (hours)	48
LMSR unload time (hours)	36
FSS (strategic) (#)	6
FSS (forward) (#)	2

(*Continued*)

Initial Deployment Parameters for the Notional Omnia Scenario (Continued)

	Value
FSS speed (knots)	28
FSS capacity (tons)	10630
FSS load time (hours)	48
FSS unload time (hours)	48
Container (stategic) (#)	5
Container (forward) (#)	0
Container speed (knots)	18
Container capacity (tons)	19643
Container load time (hours)	72
Container unload time (hours)	72
Europe RORO (strategic) (#)	0
Europe RORO(forward) (#)	10
Europe RORO speed (knots)	18
Europe RORO capacity (tons)	4800
Europe RORO load time (hours)	48
Europe RORO unload time (hours)	48
Litton HSS (strategic) (#)	0
Litton HSS (forward) (#)	0
Litton HSS speed (knots)	45
Litton HSS capacity (tons)	5000
Litton HSS load time (hours)	12
Litton HSS unload time (hours)	12
Aerocraft (strategic) (#)	0
Aerocraft (forward) (#)	0
Aerocraft speed (knots)	120
Aerocraft capacity (tons)	380
Aerocraft load time (hours)	12
Aerocraft unload time (hours)	12
Air Transportation Parameters	
Strategic air distance (nm)	6494
Through air distance (nm)	6994
Forward air distance (nm)	2822
Theater air distance (nm)	500
Number of a/c types (#)	5
C-5 (strategic) (#)	60
C-5 (through) (#)	0
C-5 (forward) (#)	0
C-5 (theater) (#)	0
C-5 speed (knots)	405
C-5 capacity (tons)	89
C-5 load time (hours)	6
C-5 unload time (hours)	2
C-17 (strategic) (#)	0
C-17 (through) (#)	20
C-17 (forward) (#)	50
C-17 (theater) (#)	0

(*Continued*)

Initial Deployment Parameters for the Notional Omnia Scenario (Continued)

	Value
C-17 speed (knots)	410
C-17 capacity (tons)	59
C-17 load time (hours)	4
C-17 unload time (hours)	2
C-130 (strategic) (#)	0
C-130 (through) (#)	0
C-130 (forward) (#)	0
C-130 (theater) (#)	150
C-130 speed (knots)	290
C-130 capacity (tons)	20
C-130 load time (hours)	2
C-130 unload time (hours)	1
CRAF (strategic) (#)	60
CRAF (through) (#)	0
CRAF(forward) (#)	0
CRAF (theater) (#)	0
CRAF speed (knots)	400
CRAF capacity (tons)	86
CRAF load time (hours)	5
CRAF unload time (hours)	3
SSTOL (strategic) (#)	0
SSTOL (through) (#)	0
SSTOL (forward) (#)	0
SSTOL (theater) (#)	0
SSTOL speed (knots)	320
SSTOL capacity (tons)	40
SSTOL load time (hours)	2
SSTOL unload time (hours)	1
Throughput Constraints	
CONUS seaport capacity (tons/day)	165000
CONUS airport capacity (tons/day)	20000
Forward S/P capacity (tons/day)	60000
Forward A/P capacity (tons/day)	14000
Strategic S/P capacity (tons/day)	18000
Strategic A/P capacity (tons/day)	5361
Theater road time (days)	1
Theater road capacity (tons/day)	8000
Theater rail time (days)	2
Theater rail capacity (tons/day)	5000
Theater water time (days)	2
Theater water capacity (tons/day)	0
Theater A/P capacity (tons/day)	3500

Lift Profile of Equipment to Move

Day	CONUS SPOE	CONUS APOE	Forward SPOE	Forward APOE	PREPO	SPOD Usage	APOD Usage	Theater Usage
1	0	0	0	0	0	0	0	0
2	0	18415	0	12924	0	0	0	0
3	0	25000	0	5000	0	0	0	0
4	40606	20000	29549	5000	62199	0	0	0
5	0	15000	0	5000	0	0	0	0
6	60000	7514	15000	0	10468	0	0	0
7	50000	5000	0	0	0	0	0	0
8	50000	0	0	0	0	0	0	0
9	30000	0	0	0	0	0	0	0
10	30000	0	0	0	0	0	0	0
11	30000	0	0	0	0	0	0	0
12	30000	0	0	0	0	0	0	0
13	30000	0	0	0	0	0	0	0

20

Reserve Manufacturing Capacity for Augmenting Contingency Logistics Requirements

Marlin U. Thomas
Air Force Institute of Technology

Mark A. Lawley
Purdue University

20.1 Introduction

American industry plays a major role in our national defense by providing the equipment, materials and supplies necessary for our military forces to maintain readiness and respond to contingencies that threaten our national security. We have always relied on the private sector to provide increased and surging logistics support for the military during wartime by shifting production plans and priorities to support war efforts. US public and private sectors have always been responsive to support our nation during times of crises and emergencies. Selected industries serve as defense partners through contracts and agreements to augment our national defense effort when called upon to manufacture war materials for contingencies. Most notable contingency manufacturing systems existed during World War II when limited war reserves made it necessary for industries to shift from normal production to produce weapons, ammunition, and supplies for supporting the war effort.

In recent years US public concerns for homeland and national security have escalated enormously, now being a top priority in our national strategy. Still, the changing economic and political conditions have altered public sentiment that once favored a major active military force to our present attempt at sustaining an all-volunteer force. This makes it necessary to place a much greater reliance on the utilization of military reserves and National Guard forces for contingency support. This also makes it necessary for an even greater reliance on civil reserve augmentation of industry as back-up for providing on-time logistics support in the event of national emergencies.

There are several versions or options for civil augmentation support for wartime production using government owned and contractor owned facilities with contracted operations. There are cases of government owned and government operated facilities such as arsenals and munitions factories that run at low levels of productivity during peacetime but carry the additional reserve capacity to surge production if needed during time of war. The primary function of this reserve augmentation is to maintain a skill pool and technology for materials that are only needed for war. Many logistics items like clothing, vehicles, and construction materials however, are found in the general market and do not require special facilities and processes for use by the military. For these cases the required reserve is to access a surge capability to meet build-up requirements for mobilizing military forces. The companies engaged in this type of reserve augmentation will generally produce other similar products but agree to serve the needs of the military as a priority source of logistics during wartime.

Industrial engineers and operations researchers have studied logistics problems for many years and there is a large literature on methods and techniques for planning logistics for manufacturing and military operations. Several classical operations research problems such as the traveling salesman, knapsack, bin-packing, trans-shipment, and Weber problems have formed the basis for algorithms and techniques for dealing with scheduling, routing, distribution, and assignment problems for achieving minimum cost operations for manufacturing and production systems under normal and stable demand conditions. Only recently have reported studies appeared on approaches and methods for dealing with contingency logistics problems. Jones (1995) described the risks and implications of material shortages for critical material needs during Major Regional Contingencies. Zografos et al. (1998) proposed an integrated framework for managing emergency response logistics operations for electrical utilities. Protecting our critical infrastructure is one of the mission critical areas for our US Homeland national security strategies. Barbarosoglu et al. (2002) proposed a model for providing helicopter support for disaster relief operations. Thomas et al. (2002) introduced a basic framework for classifying contingency logistics systems (CLS). Beamon and Kotleba (2006) examined the differences in the operational characteristics conventional commercial supply chains and those for humanitarian relief operations. In general, inventories have to be managed strategically during contingency operations due to the uncertainties and dynamics of demands. Thomas (2008) developed a time dependent model for planning inventories during sustainment operations when critical items have to be conserved. There is need for more approaches to assessing the overall capability to effectively plan for military contingencies. There is also need for guidance on increasing the preparation and readiness of civil augmented manufacturing organizations for implementing into the military logistics systems.

The purpose of this chapter is to provide a framework for the civil industrial augmentation support decision paradigm for military logistics and propose a method for examining strategies for mobilizing and programming production requirements for wartime industry augmentation. We start in Section 20.2 by describing a CLS as an element of contingency operations for military missions. Here we will restrict ourselves to those military contingencies that are major in that additional reserve resources are necessary to cope with a wartime scenario. In Section 20.3, we describe a manufacturing civil augmentation program (MANCAP) for supporting supply needs during emergency force build-up conditions. A model is presented along with an example for programming augmented production of protective clothing to meet military contingency planning requirements. Some concluding remarks are provided in Section 20.7.

20.2 Contingency Logistic Systems

A CLS is a collection of processes and procedures for providing the logistics support requirements for contingency operations. These requirements include the procurement and acquisition, material handling and packaging, transportation, storage, and distribution of the supplies, materials and equipment required for the operation. Although contingencies occur at random there are plans for most military contingencies that allow at least the initial guidelines for developing the logistics requirements and executing the actions required to support the deployment and build-up of forces. Once a contingency occurs and a response operation is initiated, the CLS planning and execution become complicated by the dynamics of the operations and prevailing risk and uncertainties throughout the life cycle. The relative resource profile for a contingency operation is shown by the ramp function in Figure 20.1. The life cycle consists of the three distinct phases: deployment, sustainment, and reconfiguration. The deployment phase starts at a time t_0 marking the beginning of the contingency and extends throughout the mobilization and force build-up to a time t_1. On location an infrastructure is developed for the operation and interface with host nations, and the service center support for food and supplies, medical, transportation, and so forth is implemented in accordance with the force size and mission. As the force builds up the logistics requirements also ramp up as shown in Figure 20.1. The slope of this ramp during (t_0, t_1) is directly related to the expenditure of resources necessary to achieve the desired military capability. The steeper the slope the greater the resource commitment or alternatively, by selecting a smaller slope the force build-up will require a longer time period. The selection of the slope and time to achieve the desired military capability are also constrained by the availability of the manpower, supplies, equipment, and transportation resources.

At time t_1 the build-up is completed and the operations continue throughout what we call the sustainment phase. During this phase the nature of the operations and logistics requirements can change with the tempo and operational conditions but the basic processes and structure are all in place. The length of the sustainment period, depending upon the contingency could be weeks, months, or years. There are numerous examples in our history that range from contingencies of short duration like the Bay of Pigs in Cuba, Operation Restore Hope for humanitarian relief of the people of Somalia, and Operation Desert Storm in Kuwait. Other contingencies led to major conflicts like Korea, Viet Nam and the world wars. Our focus here will be on those contingencies that involve significant campaigns that require major surges in logistics requirements from reserve sources.

At some point in time t_2, the operation is terminated by completion of the mission or by directive of higher authority. During this phase the infrastructure is transferred or modified for local conditions, service centers for supporting the mobilized personnel and equipment are phased down or collapsed as necessary to accommodate declining demands as redeployment occurs. Supplies and assets are redistributed and the equipment and personnel are returned.

Contingencies by their very nature involve randomness and uncertainties. There are uncertainties with the initial occurrence, the type and magnitude of the operation, and the availability and readiness of the resources. A CLS is likewise affected by these conditions and the dynamics of the activities in the operation. An effective CLS must accordingly possess the basic control and feedback capabilities that allows for flexibility and rapid systems integration to adjust to a range of shifts in demands and requirements. A CLS can be characterized by its agility, high operational tempo, capability to rapidly surge up and down, and changing decision criteria for logistics related decisions during the contingency lifecycle.

During the deployment phase of a contingency, resources are committed to achieve the mission goal of $C(t_1)$ by time t_1 in Figure 20.2. The time t_1 to reach state M will depend upon the amount of resources that are committed during the ramp-up. Two alternative paths for reaching M are shown in Figure 20.2.

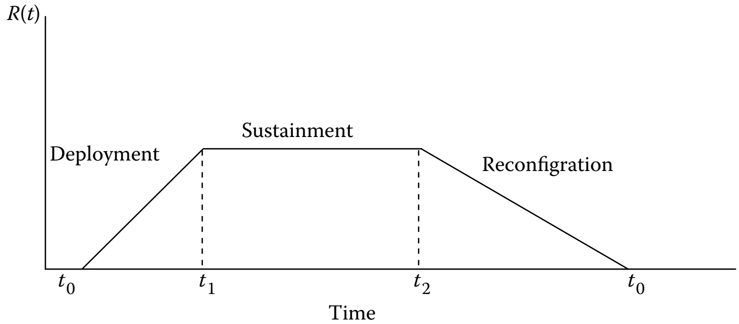


FIGURE 20.1 Lifecycle profile of a contingency logistics system.

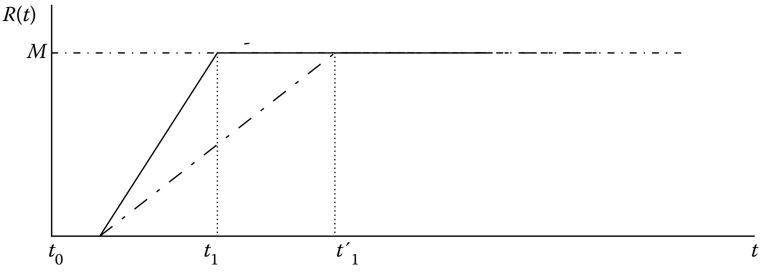


FIGURE 20.2 Alternative ramp-up programs for a contingency logistics system.

Path B takes longer and will require fewer resources per unit time than that of Path A. One or both of these alternatives can exceed existing logistics capabilities of the military and requires augmentation by reserve sources. In addition, during the sustainment phase circumstances can change and require further augmentation in order to meet logistics requirements to maintain state M.

Civil augmentation is a means for providing a reserve of critical resources to support military and contingency operations when existing capacities are exhausted. While each of the Army, Navy, Air Force, and Marine military components have reserves that are called upon for this purpose, current law restricts such recalls to Presidential order, which is not always justified nor timely to meet sporadic demands for military support that have been occurring throughout the past decade. Moreover, as we alter our policies for national security relying on all volunteer manpower and decrease the size of our military there is an even greater need to render support to national defense through reserve resources from the private sector.

During World War I and World War II manufacturers responded to the national needs and turned to support the products needed for the war efforts, which included ammunition, clothing, food, equipment, and transportation, and materials and supplies. Some of the products needed for wartime are the same product lines that serve the public during peacetime, but the demands increase during contingencies, particularly during ramp-up. For some cases, the items required are unique to the military and are not ordinarily produced in significant volumes. Examples include weapon systems, ammunitions, ordinance, protective clothing, and special chemicals and medicines. In many cases the production and manufacturing processes employed for producing these items are not so dissimilar that integrated systems could not be developed for shifting into a surge mode to produce the needed wartime products upon demand.

There were several logistics shortfalls in deploying military troops to the Persian Gulf during Operation Desert Storm in 1990. One of the earliest was an inadequate supply of desert camouflage uniforms for the thousands of personnel that were being deployed. In some cases troops were forced to advance forward wearing battle dress that not only was for the wrong climate but was also in very close resemblance to the enemy. In spite of major efforts by American manufacturers to meet the surging demands the additional capacity needed to provide the needed logistics flow of critical supplies and logistics was simply not adequate.

Civilian contractors have been utilized in supporting logistics functions for contingency and wartime military operations as far back as the Napoleonic wars. This practice continued throughout American history and the civilian support proved to be a vital arm of the military for major operations during the American Revolution, Civil War, World Wars I and II, Korea, Viet Nam and the Persian Gulf wars. The Civil Reserve Air Fleet (CRAF) is a formal program for ensuring airlift capability is available to the Department of Defense during times of emergencies when existing military capability is exceeded. Through this program selected aircraft types are identified that would be required for military contingencies, such as long-range high speed, cargo carrying, and other requirements peculiar to military needs. Airline pledges for support are contracted for various CRAF segments, which are then activated when needed for national emergencies. This was an extensive arm of DOD during the Gulf War build-up.

In 1985, the US Army formalized this requirement for engaging civilian logisticians during contingency operations through the Logistics Civil Augmentation Program (LOGCAP) (see Department of the Army 1985 and Army Material Command 1997). This program consists of contractual relationships to provide a reserve capability for logistics functions such as transportation and maintenance, fuel, ammunition, equipment, material handling, construction, medical support, and security. In December 1992, LOGCAP was used by the US Navy for Operation Restore Hope to provide humanitarian relief for the people of the Republic of Somalia. LOGCAP has since been used in other operations including Bosnia along with an off-shoot program CONCAP—Construction Civil Augmentation Program, developed by the US Naval Facilities Engineering Command for reserve construction support.

20.3 Manufacturing Civil Augmentation Program (MANCAP)

CRAF, LOGCAP, and CONCAP are more service type logistics support areas that are now integrated into the military planning processes. Except for very special items, there appears to be no similar formal arrangements for serving basic logistics needs during ramp-up and sustainment phases for major contingencies. Emergency deployment can require enormous supplies for the food, troop bed-down, medical, construction, and equipment. We propose a MANCAP that would parallel the other programs serving as a reserve resource for producing critical supplies and could be integrated into our civil and defense acquisition and procurement planning processes.

Figure 20.3 illustrates the relationship among supporting production centers for a given CLS deployment site. Unlike other civil augmentation programs, MANCAP would involve partnering and collaboration with numerous contractors and differing arrangements. These include the management and ownership relationships: (1) government-owned civilian operated (GOCO); (2) civilian owned and civilian operated (COCO); and (3) civilian owned government operated (COGO). COCO is of course the most common type of augmented support followed by GOCO, which has been used primarily for special production at arsenals. COGO is more rare but important option as we become further integrated in using civilian augment support as an integral part of our national defense system.

20.4 A Conceptual Model of Logistics Acquisition and Distribution

The logistics support for deployed military units, whether for purposes of training, exercises, or actual contingency operations, is managed and coordinated through a logistics support center. A CLS is directly linked and generally initiated through one of several centers located throughout the US for various types of support forces. For naval construction operations, the US Navy Seabees are supported by one of the US Naval Construction Battalion Centers located at Gulfport, MS and Port Hueneme, CA. The flow and

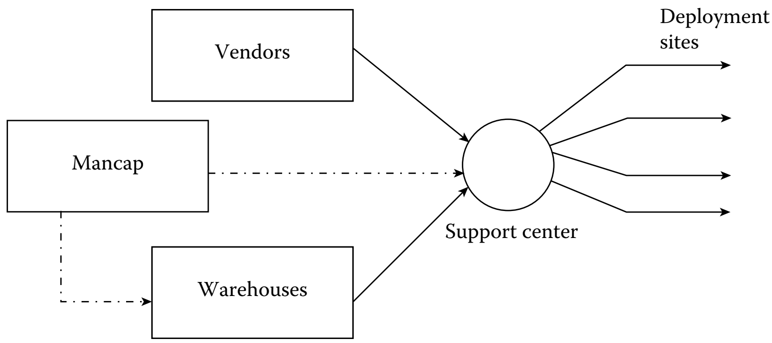


FIGURE 20.3 Procurement and distribution through a logistics support center.

distribution of supplies and materials through a logistics support center is shown in Figure 20.3. For clarity we refer to those producers in the supply chain during normal (i.e. noncontingency) conditions as vendors, and those that are elements of MANCAP as producers. It is conceivable for a given supplier to be both a vendor and a MANCAP producer. The flow of supplies from MANCAP is shown as dotted lines, indicating that it is only active during a contingency. During normal conditions the distribution center transfers supplies and equipment to various training and exercise sites where units prepare for readiness. The Center carries out the acquisition and procurement of items according to planned training and operation schedules. Items are received, prepared and packaged into kits tailored for the given operations according to the plans. When the deployment is completed, items are returned and processed for re-distribution. Kit assemblies are broken down, and damaged items are repaired or replaced and stored.

The proposed procedure for incorporating MANCAP with a CLS is to first establish the goals for the critical supply and equipment elements that will achieve the mission goal for accomplishing deployment. Critical support elements are those key items that form the critical path of a logistics network. These elements can usually be identified from a preexisting plan that can be modified to meet current requirements when it is actually executed. Based on the mission and plan, the force support requirements can be derived with key milestones defined for achieving the ramp-up level shown in Figure 20.1. There are two types of critical support elements for establishing effective logistics flow; the basic personnel support elements, and those special key items that are unique for the given mission. Standard tables of allowance are prescribed for military units that specify the requirements for food, clothing, weapons and ammunition, and basic bed-down supplies as a function of the mission. Examples of special key items include cold weather or desert organizational clothing, bottled water, water well drilling equipment, and scarce materials necessary for the mission. Once identified, the critical support requirements are converted into a logistics support network. This network forms the basis for all production, acquisition and distribution schedules. If the existing sources of supplies and equipment from warehouses and vendors are not adequate, then it becomes necessary to activate a MANCAP utilization plan.

To illustrate the logistics load for a logistics support center consider the graph in Figure 20.4. For our purposes here, $L(t)$ can be a hypothetical indicator of the deployed inventory of materials and equipment. Examples of measurements of $L(t)$ are total dollars expended, number of cargo units, and number of vehicles. In actuality, it is a vector of the key logistics elements. The first segment of the graph, from 0 to t_0 shows the typical fluctuation pattern of incoming and outgoing items that support training, exercises, and even contingencies of a minor level that require additional resources above that which is allotted for training. At time t_0 a contingency occurs which initiates the CLS. Let x_G be the level of logistics load necessary to accomplish the deployment phase of the contingency at t_1. The gradual ramp up from $L(t_0)$ to x_G at time t_1 in Figure 20.4 shows the implementation of the logistics network for the deployment and build up of forces which would include the MANCAP support if necessary.

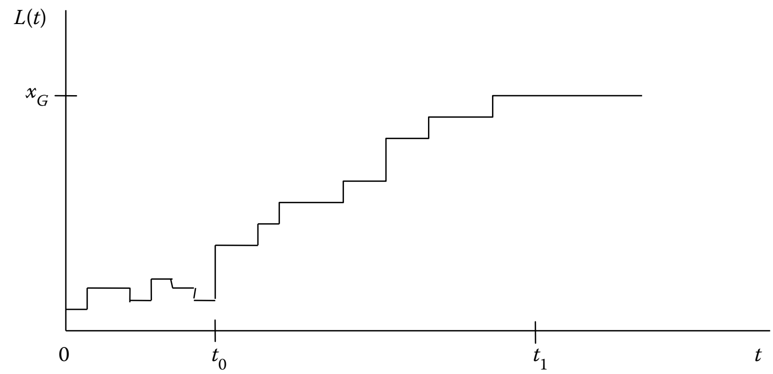


FIGURE 20.4 Load profile for a contingency.

20.5 A Model for Allocating Mancap Resources

Let us consider the problem of determining the production levels for a set of n production centers supporting the manufacturing reserve requirements. This could be viewed as an overall program for a fixed product or for a particular deployment site. So we consider a single item that can be produced through n MANCAP facilities, which we will refer to as "producers". Each producer can be considered as occupying a primary manufacturing function that produces products in the private sector but under a contractual arrangement to shift to production to meet specified demands upon activation to meet a contingency. In this contractual arrangement, each producer agrees to some range on the level of production to be dedicated to a contingency and to some time frame over which production is to be ramped up and down.

We assume that production planning will occur over a specified planning horizon, which covers some portion of the estimated contingency length, and which is partitioned into m periods, not necessarily of equal length. We further assume that total production rate at the end of each period has been determined, and that the ramp-up and ramp-down functions are linear. We shall adopt the following notation:

For $i = 1 \ldots n$ producers and $j = 1 \ldots m$ periods

Parameters:

- O_{ij}, the maximum output for producer i in period j.
- o_{ij}, the minimum output for producer i in period j.
- Θ_{ij}, the maximum ramp rate for producer i in period j.
- θ_{ij}, the minimum ramp rate for producer i in period j.
- a_{ij}, the marginal cost for producer i making one unit of product in period j.
- b_{ij}, the marginal cost for producer i ramping up or down one unit in period j.
- P_j, the target production level at the end of period j.
- x_{i0}, the production level of producer i at time 0.
- t_j, length of period j.

Variables:

- x_{ij}, production level of producer i at the end of period j ($\in \Re^+$).
- y_{ij}, ramping rate for producer i in period j ($\in \Re$).

Note that O_{ij}, o_{ij}, Θ_{ij}, θ_{ij}, a_{ij}, and b_{ij} capture the contractual arrangements agreed to by the producers. The P_j reflect the time phased product demand generated by the contingency, while the x_{io} capture the production levels of the producers at the onset of the contingency. The x_{ij} variables represent the individual production rates that producers will be asked to achieve by the end of each production period, while the y_{ij} represent the rates at which the producers will be asked to ramp their production rates during these periods. (We note that these two variables are not independent and that the problem can be easily formulated with either alone. We include both in the following formulation for the sake of clarity.)

We now model the problem of determining the time phased production levels for the set of n producers as follows:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} \left(a_{ij} x_{ij} + b_{ij} y_{ij} \right)$$

st.

$$\sum_{i=1}^{n} x_{ij} = P_j \quad j = 1, \ldots, m$$

$$
\begin{aligned}
&o_{ij} \le x_{ij} \le O_{ij} \quad i=1\ldots n,\ j=1\ldots m\\
&\theta_{ij} \le y_{ij} \le \Theta_{ij} \quad i=1\ldots n,\ j=1\ldots m\\
&x_{ij} = x_{1,j-1} + y_{ij}t_j \quad i=1\ldots n,\ j=1\ldots m\\
&x_{ij} \in R^+ \quad \text{and} \quad y_{ij} \in R \quad i=1\ldots n,\ j=1\ldots m
\end{aligned}
\tag{20.2}
$$

This simple formulation minimizes the production and ramping costs over the planning horizon while assuring that desired production rates are achieved and that no producer is asked to accelerate production too rapidly. Furthermore, the model provides useful support in developing and managing MANCAP contracts, since the shadow prices of the dual can be used to help develop the allocation of contractual requirements among producers.

20.6 An Example for Personal Protective Clothing

Consider the situation in which five MANCAP facilities have been developed for supplying individual protective clothing for chemical, biological and radiological (CBR) exposure. An individual outfit consists of a hooded jacket, trousers, boots, gloves, and face mask. In developing the program each supplier was selected based on available production capacity, costs, and capabilities for producing high quality supplies for anticipated contingencies. The suppliers were contracted as follows:

S_1,S_2: Provide rapid transition to produce new demands at higher cost;
S_4,S_5: Support production over longer term at lower cost but require longer ramp period;
S_3: Cover the mid-range period between activation of short and long term suppliers.

The contingency plan calls for the following three time periods and CBR demand levels for building up a force structure:

Period j	1	2	3
Duration (days)	0,30	31, 60	61,150
Units demanded, P_j	20,000	50,000	150,000

The production parameters specified in the contractual arrangements for each of the suppliers are given in Table 20.1.

TABLE 20.1 Parameters for the Personal Protective Clothing Example

Producer	Maximum Production θ_i (units)	Maximum Ramp-up Θ_i (units/day)	Period j	Unit Cost ($/unit)
S_1	50,000	6000	1	1000
S_2	50,000	6000	1	1000
S_3	75,000	5000	2	500
S_4	100,000	2000	2	100
		8000	3	
S_5	100,000	2000	2	50
		8000	3	

20.6.1 Solution

The problem is to determine time phased output levels for the five MANCAP production facilities that will meet the given contingency demands for CBR clothing. Applying a conventional method the linear program of Equation 20.1 can be solved to obtain the solution given in Table 20.2. Producers S_1 and S_2 would be augmented immediately to provide $x_{11}=x_{21}=10{,}000$ units during the initial 30 day period. Producer S_3 would then be augmented to produce $x_{32}=26{,}000$ units in Period 2 along with S_4 and S_5 that would provide $x_{42}=x_{52}=12{,}000$ units. After 60 days, Producers S_4 and S_5 would support the mission by providing $x_{43}=100{,}000$ and $x_{53}=50{,}000$ CBR units over the remaining Period 3. The production profile for this solution is shown in Figure 20.5.

20.7 Summary and Concluding Remarks

Contingencies have always existed and they are inevitable for our future. In this chapter our focus has been on a framework for incorporating civil augmentation into the pool of resources for supporting CLS. The models presented can be useful in making planning decisions but there is much to be done to fill in this framework.

The role of logistics is crucial for the effectiveness of any operations whether for military, business, or public purposes. The analysis of logistics support systems is difficult even under stable conditions due to the number and complexities of decisions involved in carrying out the support activities. Formal guidance exists for analyzing defense logistics support under general conditions (Department of Defense 1997). However, these procedures do not cover integrating civil augmentation as part of the process.

While the concept for MANCAP is relatively simple, several methods and processes are needed for its implementation. First, there are design and development issues in constructing an effective augmentation program. What are the critical logistics items that should be included in the program? Perishable items, such as drugs and chemical products, specialized clothing, munitions, life support and building

TABLE 20.2 MANCAP Production Schedule for CBR Unit

Producer *i*	Period *j*		
	1	2	3
S_1	10,000	0	0
S_2	10,000	0	0
S_3	0	26,000	0
S_4	0	12,000	100,000
S_5	0	12,000	50,000

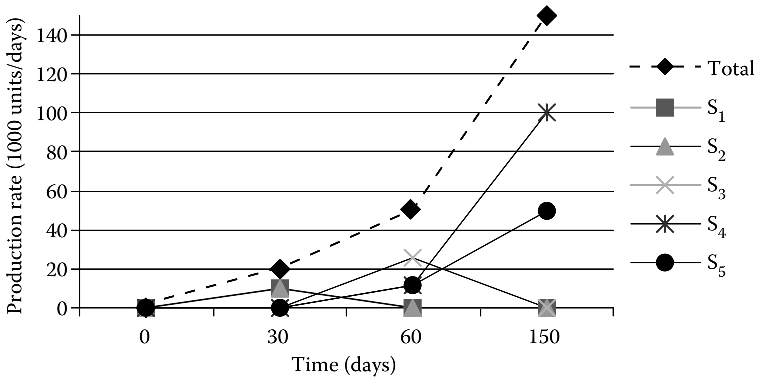


FIGURE 20.5 Production profiles for the CBR example.

materials are likely candidates. This will depend on the volume, inventory and storage requirements, and factors pertaining directly to the criticality of the particular mission plans. A related issue is how many and what type of producers should be in the program? The allocation model described in Section 20.5 provides some high level guidance for determining alternative allocations, given a fixed number of producers. In addition, decisions on the number of producers necessary and how to establish their relative value are important in constructing a program. Contracting is another issue that is important in establishing cost and timely effective support for the system. LOGCAP is based on cost plus contracts awarded typically to a single company. MANCAP will require contracting arrangements with different producers and various product lines. There are important socio-economic issues that arise in establishing the cost and value of contract support under the different products, planning scenarios, and risk levels that are to be included in our national security and defense planning.

There are also issues in managing CLS operations. How much of the inventory should be provided through direct procurement, warehouse inventories, and reserve augmentation? Once a contingency is initiated and forces are deployed, then it is necessary to establish the acquisition and procurement needs to support the dynamics of the operations. In addition there is the question of, how should the production schedules be determined and allocations made among the producers? There is need for models to support these types of decisions.

The purpose for civil augmentation is to take advantage of the resources available in private industry to fulfill the surging skills and additional manufacturing capacity needed by our military in event of major contingencies. A program such as MANCAP should therefore be integrated with our military strategy and be included in the overall readiness for national security. Accordingly, questions need to be addressed like what are the measures of effectiveness and readiness that can be used for defense planning? Also, how can the MANCAP program be exercised to ensure compliance with a reliable contingency and wartime support capability?

These issues establish a set of future needs for incorporating MANCAP in contingency plans and operations. Realistically, we must rely on civilian augmentation to fulfill and afford the homeland and national security needs of our nation.

References

Barbarosoglu, G., Ozdamar, L., and Cevik. A. 2002. An interactive approach for hierarchical analysis of helicopter logistics in disaster relief operations. *European Journal of Operational Research*, 140, 118–133.

Beamon, B. and Kotleba, S. 2006. Inventory modeling for complex emergencies in humanitarian relief operations. *International Journal of Logistics*, 9(1), 1–18.

Department of the Army. 1985. AR 700-137, Logistics Civil Augmentation Program (LOGCAP), Washington, DC.

Army Material Command. 1987. AMC Pamphlet No. 700-30, Logistics Civil Augmentation Program, Alexandria, VA.

Department of Defense. 1997. MIL-HDBK-502, Acquisition Logistics Handbook. Defense Logistics Agency, Washington, DC.

NRC (National Research Council). 2002. *Munitions Manufacturing—A Call for Modernization*. National Academy Press, Washington DC.

Thomas, M.U., Deshpande, V., and Iyer, A.V. 2002. Contingency Logistics Systems—A Taxonomy and Application Modeling Framework for Contingency Logistics Planning. Tech. Memo. No. 02-03, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

Thomas, M.U. 2008. Inventory models for contingency operations In *Handbook of Military Industrial Engineering*, A.B. Badiru and M.U. Thomas (Eds), 21-1–21-10. Taylor & Francis, Baco Raton, FL.

Zografos, K., Douligeris, C., and Tsoumpas, P. 1998. An integrated framework for managing emergency-response logistics: The case of the electric utility companies. *IEEE Transactions on Engineering Management*, 45(2), 115–126.

21

Inventory Models for Contingency Operations*

Marlin U. Thomas
Air Force Insitute of Technology

21.1 Introduction

The public awareness and threat of upheaval from terrorist attacks, natural disasters, and other uncertain crises events has increased dramatically over the past two decades. Unfortunately, we have to position ourselves to deal with a broad range of contingency events that threaten the safety and security of innocent victims within our country and throughout the world.

Contingencies arise from crises events that require emergency action to assist and protect people from disastrous situations. They are not restricted to military action, though often military support is provided for disasters occurring within the continental U.S. as well as abroad. Essentially all contingency operations require significant logistics functions for providing procurement, storage, transportation, and distribution of supplies and materials for meeting rigorous time and scheduling requirements. While these logistics functions are rather standard for any operational system, they are further complicated in contingency operations due to the prevailing uncertainties and dynamics of demands and mission times. This is particularly crucial for critical inventories that must be managed strategically to ensure mission success. Very little work has been reported that deals with these conditions. Thomas et al. (2002) introduced a basic framework for classifying contingency logistics systems. Beamon (2004) compared the operational characteristics of supply chains for humanitarian relief operations in comparison to conventional commercial supply chains and proposed a set of performance measures for the contingency operations.

This chapter presents some models for establishing inventories for critical resources that are necessary to sustain operations during a contingency. We start in Section 21.2 with a description and framework for contingency logistics systems (CLS). Every CLS has three operational phases: deployment, sustainment, and reconfiguration. In Section 21.3, we formulate the inventory decision problem for

* This chapter is a modification of "Inventory support decisions for contingency operations," that appeared in the *Proceedings of the Industrial Engineering Research Conference*, May 2004.

sustained operational conditions. Results are provided in Section 21.4 for the case of a single critical item with Poisson distributed demand and random mission time. The details of the derivations are provided in Appendix A. Section 21.5 provides some concluding remarks and suggested directions for further modeling needs.

21.2 Contingency Logistics Systems

A CLS is defined as a collection of processes and procedures for providing the procurement and acquisition, material handling and packaging, transportation, storage and distribution of the supplies, materials and equipment required for contingency operations (Thomas et al., 2002). While contingencies are driven by random events, plans exist for most military and civil contingencies that allow at least the initial guidelines for developing the logistics requirements and execution actions to support the deployment and build-up of a force. Every contingency has three distinct phases as shown in the ramp profile of Figure 21.1. $R(t)$ represents the logistics resources employed as a function of time t into the deployment. This could be measured in total dollars expended, number of cargo units, personnel, number of vehicles, or some weighted combination of essential logistics. The deployment phase starts at time t_0 which marks the beginning of the contingency and it extends throughout the mobilization and organizational build-up of the force at a time t_1. Once the emerging force reaches its site or location, an infrastructure is developed for the operation and service center support for food, supplies, medical support, transportation and so forth are implemented as appropriate for the force size and mission. As the force builds up over time, the logistics requirements $R(t)$ also ramp up as shown in Figure 21.1. The rate of build-up, i.e. the slope t_1, t_2 will depend upon the resources expended over time. Though the cost of the operation is important this is not necessarily the primary critera for decision making during the deployment phase for contingencies.

At time t_1 the build-up is completed and the operations continue throughout what we call the Sustainment Phase. During this phase the nature of the operations and logistics requirements can change with the tempo and operational conditions but the basic processes and structure are all in place. The time to reach t_1 will generally depend on the resources expended to get there. This will vary considerably with the type of mission. Contingency plans provide for allotted inventories to be available for emergencies. Once operations start, resources are regulated to cope with the mission requirements; but uncertainties exist relative to the dynamics of ongoing operations and uncertainty of duration. The length of the sustainment period, depending upon the contingency could be weeks, months, or years. During this phase there is a high degree of uncertainty about the mission and logistics requirements and the strategy is then to conserve resources to the extent possible. Critical items like ammunition, petroleum products, medical supplies, and even food products can have a major impact on the mission. Inventory decisions under these circumstances can often be treated as decisions under risk,

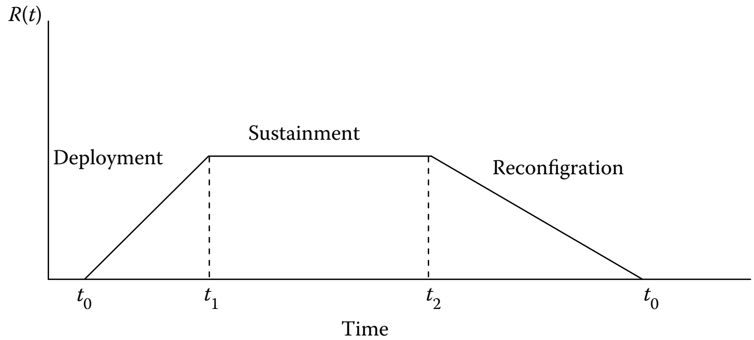


FIGURE 21.1 Lifecycle profile of a contingency logistics system.

by assuming distributions about the random demands and mission times. At some point in time t_2, the operation is terminated by completion of the mission or by directive of higher authority. During this phase the infrastructure is transferred or modified for local conditions, service centers for supporting the mobilized personnel and equipment are phased down or collapsed as necessary to accommodate declining demands as redeployment occurs. Supplies and assets are redistributed and the equipment and personnel are returned.

Contingencies by their very nature involve randomness and uncertainties. There are uncertainties with the initial occurrence, the type and magnitude of the operation, and the availability and readiness of the resources. A CLS is likewise affected by these conditions, and the dynamics of the activities in the operation.

21.3 Time Dependent Inventory Decisions

In this chapter we consider a hierarchical system consisting of a distribution center that receives and distributes the supplies and materials to subordinate elements in the chain. For convenience, each subordinate element can be considered a regional operational element of the overall system. For military missions this could be a higher command with subordinate commands providing various operational and support functions for different geographic sectors. So the challenge is to establish the amount of inventory to maintain for the system of subordinates for the duration of the period. Ordinarily this could be achieved through a simple single stage inventory model but during contingency operations there are many uncertainties that enter decision problems.

21.3.1 Assumptions and Notation

We consider a single inventory item that is critical to the mission. The demand is represented by a discrete random variable X defined on $\mathbb{N}$ and the mission time T is a continuous random variable defined on $\mathbb{R}_+$. We shall assume that the initial inventory provision is to cover all units for the total mission. The notation is as follows:

c_1	unit cost of inventory overstocked in \$/unit.
c_2	unit cost of inventory under stocked in \$/unit.
I	total mission inventory in units.
λ	item demand rate in units/time.
X	a random variable representing demand in number of units.
T	random variable representing mission time in
$P_{X\|T}(x)$	conditional distribution of demand X given mission time T
t_b	constant lower bound on mission time.
$G_T(t)$	distribution of mission time T.
$H_{X,T}(x,t)$	joint distribution of demand and mission time.

We also assume that $c_2 >> c_1$, the cost of being under-stocked is very much greater than that of being overstocked.

21.4 Single Period Models with Risk

Incorporating a mission time dependency with the single stage inventory model the expected total cost for an inventory level of I is given by

$$C(I,t) = c_1 \sum_{x=0}^{I} (I-x)h(x,t) - c_2 \sum_{x=I+1}^{\infty} (x-I)h(x,t), \quad x = 0,1,\ldots;\ t \geq 0\,, \tag{21.1}$$

where

$$h(x,t) = P(X = x,\, T \le t) = \int_0^t P_{X|T}(x)\,dG_T(\tau) \tag{21.2}$$

is the joint discrete probability mass function for the demand at x and mission time of t, and the joint cumulative distribution function is given by

$$H_{X,T}(x,t) = P\{X \le x, T \le t\} = \sum_{k=0}^{x} h(k,t) \tag{21.3}$$

Using the minimum expected cost decision criteria, the optimum inventory level is found by solving Equation 21.1. This is a discrete optimization problem with $C(I,t)$ being a convex function in x. The optimum inventory level for the distribution center is therefore the value I^* that satisfies the conditions:

$$\begin{aligned} &C(I^*+1,t) - C(I,t) \ge 0 \\ &C(I^*-1,t) - C(I,t) \ge 0 \end{aligned} \tag{21.4}$$

It follows that by substituting for I+1, I, and I–1 from Equation 21.1 and simplifying; for a given $t \ge 0$,

$$H_{X,T}(I^*-1,t) \le \frac{c_2}{c_1 + c_2} \le H_{X,T}(I^*+1,t) \tag{21.5}$$

21.4.1 Poisson Demand with Mission Time Distributed Uniform

Let the demand X be distributed Poisson with the rate $\lambda > 0$. Though the actual demand distribution for contingency conditions is essentially never known, the Poisson assumption is common for planning purposes since the corresponding inter-event times are distributed exponential which is highly random, thus making the assumption on the conservative side. Making a further conservative assumption, suppose the mission time is known only to be somewhere uniformly distributed over the interval (0,a). Substituting the Poisson probability mass and uniform probability density functions into Equation 21.2,

$$h(x,t) = \frac{1}{a}\int_0^t \frac{(\lambda\tau)^x}{x!} e^{-\lambda\tau} d\tau \tag{21.6}$$

It follows that (see Proposition 1, Appendix A),

$$h(x,t) = \left(\frac{1}{a}\right) h(x-1,t) - \left(\frac{1}{\lambda a}\right)\frac{(\lambda\tau)^x}{x!} e^{-\lambda\tau}, \quad x = 1,2,\ldots;\ t \ge 0 \tag{21.7}$$

and further,

$$h(x,t) = \left(\frac{1}{a}\right)^x h(0,t) - \frac{1}{\lambda a}\sum_{j=1}^{x}\left(\frac{1}{a}\right)^{x-j}\frac{(\lambda t)^j}{j!} e^{-\lambda t}, \tag{21.8}$$

where

$$h(0,t) = \frac{1 - e^{-\lambda t}}{\lambda a}.$$

Therefore, substituting into Equation 21.7.

$$h(x,t) = \frac{1}{\lambda a}\left\{\left(\frac{1}{a}\right)^x \left(1 - e^{-\lambda t}\right) - \sum_{j=1}^{x}\left(\frac{1}{a}\right)^{x-j}\frac{(\lambda t)^j}{j!}e^{-\lambda t}\right\} \tag{21.9}$$

from which it follows that

$$H(x,t) = \frac{1}{\lambda a}\left\{\left(1 - e^{-\lambda t}\right)\frac{a^{x+1} - 1}{a^x (a-1)} - \sum_{k=1}^{x}\sum_{j=1}^{k} a^{k-j}\frac{(\lambda t)^j}{j!}e^{-\lambda t}\right\}, \quad x = 0,1,\ldots;\ 0 < t \le a. \tag{21.10}$$

Example 21.1

The contingency planning process involves lots of assumptions, scenario developments, and analyses and evaluations. The objective is to establish a preliminary set of conditions that can be used for planning the overall operations and supporting logistics requirements to ensure mission success. As an example, consider a contingency for which the anticipated mission will be of an average duration of 15 days during which the demand for a critical item is Poisson distributed at a rate of 1 per unit of time. We will assume that the nature of the crises is such that the deployment activities can be ignored, at least for this analysis. The item costs \$5,000, with an additional charge of \$15,000 for emergency procurement.

The parameter value for the Poisson demand rate is $\lambda = 1$ per unit of time, and for the maximum time $a = 30$ days (since $E(T) = a/2$). Thus substituting these values into Equation 21.10,

$$H(x,t) = \frac{1}{30}\left(1 - e^{-t}\right)\frac{30^{x+1} - 1}{30^x (29)} - \sum_{k=1}^{x}\sum_{j=1}^{k} 30^{k-j}\frac{t^j}{j!}e^{-t}, \quad x = 0,1,\ldots;\ 0 < t \le 30$$

The cost parameters are $c_1 = \$5{,}000$ and $c_2 = \$15{,}000$, therefore $c_2/(c_1 + c_2) = 0.75$ and in Equation 21.5,

$$H(I^* - 1, t) \le 0.75 \le H(I^* + 1, t) \tag{21.11}$$

If we truly believe in our assumptions and estimates then it is only a matter of computing the value of I^* that satisfies Equation 21.11. Computations for $H(x, t)$ over $x = 0, 1, \ldots, 40$ for values of $t = 23$ and 25 days are shown in the graphs of Figure 21.2. The points of intersection of these curves with the dotted line at $H(x,t) = 0.75$ show that

$$H(23,25) = 0.7495 < 0.75 < 0.7671 = H(24,25)$$

and,

$$H(26,23) = 0.7475 < 0.75 < 0.7538 = H(27,23).$$

Further calculations confirm that for a given time t the inventory level is relatively robust. As the cost of shortages in the field gets larger relative to the unit cost, the need for increasing the inventory level of

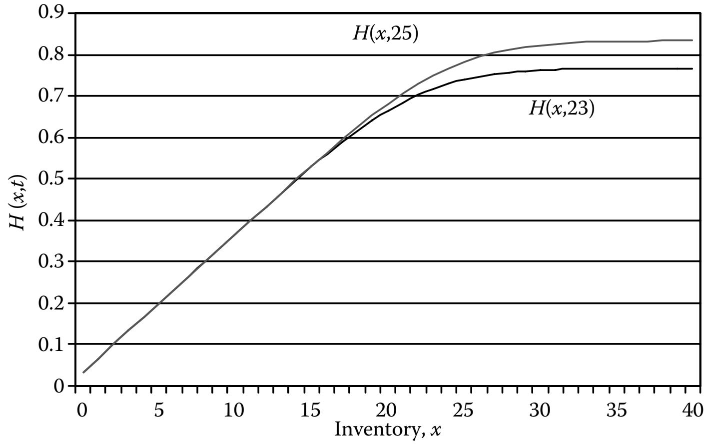


FIGURE 21.2 $H(x,23)$ and $H(x,25)$ versus inventory level for Example 1.

course becomes more critical. For the specified condition of $c_2/(c_1+c_2)=0.75$ the computed range of 23–25 units would seem to be adequate for planning purpose unless there was additional information that would add more optimism for an earlier completion time or lower value for the cost of running out of stock during the sustainment operations.

21.4.2 Poisson Demand and Exponential Distributed Mission Time

Let the demand X be distributed Poisson with rate $\lambda>0$ and suppose the mission time is distributed exponential with parameter $\mu>0$. Substituting the Poisson probability mass and exponential probability density functions into Equation 21.2 leads to the joint mixed probability density and mass function (see Proposition 3, Appendix A),

$$h(x,t)=\frac{\lambda}{\lambda+\mu}h(x-1,t)-\frac{\mu}{\lambda+\mu}\frac{(\lambda t)^x}{x!}e^{-(\lambda+\mu)t},\quad x\geq 1,\ t\geq 0 \tag{21.12}$$

where

$$h(0,t)=\frac{\mu}{\lambda+\mu}\left(1-e^{-(\lambda+\mu)t}\right).$$

The cumulative distribution is then computed by summing through Equation 21.11 to obtain the following

$$H(x,t)=\left(1-e^{-(\lambda+\mu)t}\right)\left[1-\left(\frac{\lambda}{\lambda+\mu}\right)^{x+1}\right]-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}\sum_{k=1}^{x}\sum_{j=1}^{k}\left(\frac{\lambda}{\lambda+\mu}\right)^{k-j}\frac{(\lambda t)^j}{j!},\quad x\geq 0,\ t\geq 0. \tag{21.13}$$

21.4.2.1 Unconditional Mission Time

We note that as $t\to\infty$ in Equation 21.13 the exponential term $e^{-(\lambda+\mu)t}\to 0$ thus giving the unconditional demand distribution

$$H(x)=1-\left(\frac{\lambda}{\lambda+\mu}\right)^{x+1},\quad x\geq 0 \tag{21.14}$$

which is geometric. Further, from Equation 21.5 it follows that the optimum inventory quantity is given by

$$I^* \simeq \ln\left(\frac{c_1}{c_1+c_2}\right) \Big/ \ln\left(\frac{\lambda}{\lambda+\mu}\right). \tag{21.15}$$

This result applies when the risk of mission time is represented by the conditional distribution and the demand distribution is computed by "unconditioning," or simply integrating over all values of $t \geq 0$.

21.4.2.2 Conditioned Mission Time

Now consider the case where there is reason to believe that the mission will last at least a time t_b, but there is no reason to assume any further information about T. We therefore define the event, $B = \{T \geq t_b\}$ and proceeding from Equation 21.3.

$$H(x,B) = \sum_{k=0}^{x} \int_B \mu \frac{(\lambda\tau)^k}{k!} e^{-(\lambda+\mu)\tau} d\tau \tag{21.16}$$

Integrating by parts and simplifying, the probability mass function is given by

$$h(x,B) = \left(\frac{\mu}{\lambda+\mu}\right)\frac{(\lambda t_b)^x}{x!} e^{-(\lambda+\mu)t_b} + \left(\frac{\lambda}{\lambda+\mu}\right) h(x-1,B), \quad x \geq 1, \tag{21.17}$$

where

$$h(0,B) = \left(\frac{\mu}{\lambda+\mu}\right) e^{-(\lambda+\mu)t_b}.$$

Example 21.2

To further illustrate the effect of mission time risk on the inventory decision let us consider an example where the demand rate is $\lambda = 1$ and the mean mission time is 30 days, or $\mu = 1/30$, but the mission time is expected to be no less than $t_b = 5$ days. Suppose the cost of being short of this item is five times the cost of being over stocked, therefore

$$\frac{c_1}{c_1+c_2} = 0.1667$$

Starting with the case of unconditional mission time from Equation 21.13 we have

$$H(x) = 1 - \left(\frac{30}{31}\right)^{x+1}, \quad x \geq 0 \tag{21.18}$$

and from Equation 21.14.

$$I^* \simeq \ln(0.1667) \Big/ \ln\left(\frac{1}{1+(1/30)}\right) = 55$$

Now if we include the stipulation that the mission will likely be at least $t_b = 5$ days in duration, then from Equation 21.11 the joint probability mass function is given by

$$h(x,B) = (1.84 \times 10^{-4})\frac{5^x}{x!} + 0.9677 h(x-1,B), \quad x \geq 1 \tag{21.19}$$

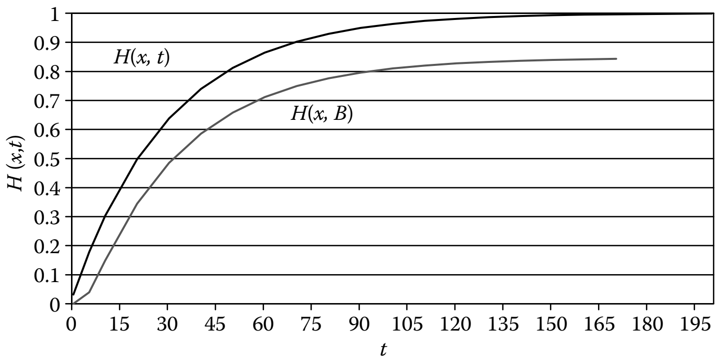


FIGURE 21.3 $H(x, t)$ versus t for Poisson demand and exponential distributed mission time.

with

$$h(0,B) = 1.84 \times 10^{-4}.$$

Summing over Equation 21.19 we find the value of the joint cumulative satisfying Equation 21.5 leads to the result of $I^* \simeq 125$ units, which is considerably larger than the 55 units when there is no assumption about the length of the mission other than through its probability density function. The effect of the assumed mission time event B on the demand distribution is shown in Figure 21.3.

21.5 Concluding Remarks

This chapter has provided results for supporting critical procurement and allocation decisions for contingency operations during the Sustainment Phase where the length of engagement is not known and conservative measures are necessary. Logistics functions have always been major factors in planning and executing contingency operations. Though many models have been developed for inventory decisions, the common assumptions about infinite planning horizons are not adequate for contingency plans. Risk and uncertainty are always present in contingency decisions. The Poisson assumption used in this chapter is considered reasonable in view of the lack of memory property with the associated interevent times. As further information is available regarding a given mission additional features can be incorporated into the mission time considerations. There are three options for incorporating mission time dependencies in the model presented in Section 21.4. The first is to assume a known distribution for the mission time G_T as was done in this chapter. A second option is to treat the parameter for the demand Λ, as a random variable with some known distribution and then compute the joint distribution $H(.)$ and corresponding conditions from Equations 21.3 and 21.4. Both of these options require an assumed distribution. For cases where this is not possible, though some factual information is available and can be appropriately structured then a maximum entropy distribution can be derived for $P_{X|T}$ in Equation 21.2. Maximum entropy concepts have been applied successfully in many such applications and are thought to be a rationale basis for making decisions when the amount of information is limited (Thomas, 1979). Hopefully, the results in this chapter will stimulate further thoughts on dealing with the mission time dynamics and uncertainties in making critical inventory decisions for contingency logistics systems.

There is need for further modeling efforts that are specific to contingency logistics systems and related operational problems. Beamon and kotleba (2006) discuss the problem and unique characteristics of modelling inventory to support humanitarian relief operations. Thomas and Lawley (2003) proposed a manufacturing civil augmentation program (MANCAP) for incorporating reserve manufacturing capacity for private industry to handle surge requirements for major contingency deployments.

The whole area of readiness planning that includes contingency logistics is in need of strategies and methods for planning rapid deployment and effective sustainment operations. Thomas (2002) proposed using reliability as a measure for quantifying the effectiveness of supply chains associated with contingency operations. Other measures and methods are needed as well.

Acknowledgments

This work was made possible through the support of the National Science Foundation, Grant 0071935-DMI. The author is also grateful for the valuable comments and suggestions of one of the referees who helped improve the manuscript.

References

Beamon, B.M. 2004. Humanitarian relief chains: Issues and challenges. *34th International Conference on Computers and Industrial Engineering*, San Francisco, CA.

Beamon, B. and S. Kotleba. 2006. Inventory modeling for complex emergencies in humanitarian relief operations. *International Journal of Logistics*, 9(1), 1–18.

Thomas, M.U. 2002. Supply chain reliability for contingency operations. *Proceedings Annual Reliability and Maintainability Symposium,* Seattle, WA, 61–67.

Thomas, M.U. and M.A. Lawley. 2003. Manufacturing for contingency logistics. Tech. Memo No. 03-05, School of Industrial Engineering, Purdue University, West Lafayette, IN.

Thomas, M.U., V. Deshpande, and A.V. Iyer. 2002. Contingency Logistics Systems—A Taxonomy and Application Modeling Framework for Contingency Logistics Planning. Tech. Memo No. 02-03, School of Industrial Engineering, Purdue University, West Lafayette, IN.

Thomas, M.U. 1979. A generalized maximum entropy principle. *Operations Research*, 27(6), 1196–1198.

Appendix A

1. Given the demand X is distributed Poisson with rate $\lambda > 0$ and the mission time T is distributed $U(0, a)$, in Equation 21.6

$$h(x,t) = \frac{1}{a}\int_0^t \frac{(\lambda\tau)^x}{x!} e^{-\lambda\tau} d\tau, \quad x = 0,1,2,\ldots; \ t \geq 0$$

Proposition 21.1

$$h(x,t) = \left(\frac{1}{a}\right) h(x-1,t) - \left(\frac{1}{\lambda a}\right)\frac{(\lambda\tau)^x}{x!} e^{-\lambda\tau}, \quad x = 0,1,2,\ldots; \ t \geq 0$$

Proof:

Integrating by parts in Equation 21.6 with

$$u = \frac{(\lambda\tau)^x}{x!} \quad \text{and} \quad dv = e^{-\lambda\tau} d\tau$$

then

$$du = \frac{\lambda(\lambda\tau)^x}{(x-1)!} d\tau, \quad \text{and} \quad v = \frac{e^{-\lambda\tau}}{-\lambda}$$

thus,

$$h(x,t) = \left[uv - \int v\,du\right]_0^t = \frac{1}{a}\left[\frac{(\lambda t)^x}{x!}\left(\frac{e^{-\lambda t}}{-\lambda}\right) - \int_0^t \left(\frac{e^{-\lambda\tau}}{-\lambda}\right)\frac{\lambda(\lambda\tau)^{x-1}}{(x-1)!} d\tau\right]$$

which simplifies to Equation 21.7

Proposition 21.2

In Equation 21.8,

$$h(x,t) = \left(\frac{1}{a}\right)^x h(0,t) - \frac{1}{\lambda a}\sum_{j=1}^{x}\left(\frac{1}{a}\right)^{x-j}\frac{(\lambda t)^j}{j!} e^{-\lambda t},$$

$$h(0,t) = \frac{1-e^{-\lambda t}}{\lambda a}.$$

Proof:

By induction,

$$h(1,t) = \left(\frac{1}{a}\right) h(0,t) - \frac{1}{\lambda a}(\lambda t) e^{-\lambda t}$$

$$h(2,t) = \frac{1}{a} h(1,t) - \frac{1}{\lambda a}\frac{(\lambda t)^2}{2!} e^{-\lambda t} = \frac{1}{a}\left\{\left(\frac{1}{a}\right) h(0,t) - \frac{1}{\lambda a}(\lambda t) e^{-\lambda t}\right\} - \frac{1}{\lambda a}\frac{(\lambda t)^2}{2!} e^{-\lambda t}$$

$$= \left(\frac{1}{a}\right)^2 h(0,t) - \left(\frac{1}{\lambda a}\right)(\lambda t) e^{-\lambda t} - \left(\frac{1}{\lambda a}\right)\frac{(\lambda t)^2}{2!} e^{-\lambda t}$$

$$h(3,t)=\frac{1}{a}h(2,t)-\left(\frac{1}{\lambda a}\right)\frac{(\lambda t)^3}{3!}=\frac{1}{a}\left\{\left(\frac{1}{a}\right)^2 h(0,t)-\left(\frac{1}{\lambda a}\right)(\lambda t)e^{-\lambda t}-\left(\frac{1}{\lambda a}\right)\frac{(\lambda t)^2}{2!}e^{-\lambda t}\right\}$$

$$-\left(\frac{1}{\lambda a}\right)\frac{(\lambda t)^3}{3!}$$

$$=\left(\frac{1}{a}\right)^3 h(0,t)-\left(\frac{1}{\lambda a}\right)\left(\frac{1}{a}\right)^2(\lambda t)e^{-\lambda t}-\left(\frac{1}{\lambda a}\right)\left(\frac{1}{a}\right)\frac{(\lambda t)^2}{2!}e^{-\lambda t}$$

…

$$h(x,t)=\left(\frac{1}{a}\right)^x h(0,t)-\left(\frac{1}{\lambda a}\right)\sum_{j=1}^{x}\left(\frac{1}{a}\right)^{x-j}\frac{(\lambda t)^j}{j!}e^{-\lambda t}$$

2. Let the demand X be distributed Poisson with rate $\lambda>0$ and the mission time T is distributed exponential with parameter $\mu>0$.

Proposition 21.3

In Equation 21.12,

$$h(x,t)=\frac{\lambda}{\lambda+\mu}h(x-1,t)-\frac{\mu}{\lambda+\mu}\frac{(\lambda t)^x}{x!}e^{-(\lambda+\mu)t},\quad x=1,2,\ldots;\ t\geq 0$$

$$h(0,t)=\frac{\mu}{\lambda+\mu}\left(1-e^{-(\lambda+\mu)}\right)$$

Proof:

Substituting in Equation 21.2 for $P_{X|T}$ and G_T,

$$h(x,t)=\mu\int_0^t\frac{(\lambda\tau)^x}{x!}e^{-(\lambda+\mu)\tau}d\tau$$

Integrating by parts with

$$u=\frac{(\lambda\tau)^x}{x!}\quad\text{and}\quad dv=\mu e^{-(\lambda+\mu)\tau}d\tau$$

thus,

$$du=\frac{\lambda(\lambda\tau)^x}{(x-1)!}d\tau,\quad v=\frac{\mu e^{-(\lambda+\mu)t}}{-(\lambda+\mu)}$$

and

$$\left[uv-\int vdu\right]_0^t=\frac{\lambda}{\lambda+\mu}\int_0^t\mu\frac{(\lambda\tau)^{x-1}}{(x-1)!}e^{-(\lambda+\mu)\tau}d\tau-\frac{\mu}{\lambda+\mu}\frac{(\lambda t)^x}{x!}e^{-(\lambda+\mu)t}$$

from which it follows that

$$h(x,t)=\frac{\lambda}{\lambda+\mu}h(x-1,t)-\frac{\mu}{\lambda+\mu}\frac{(\lambda t)^x}{x!}e^{-(\lambda+\mu)t}.$$

Proposition 21.4

In Equation 21.13,

$$H(x,t)=\left(1-e^{-(\lambda+\mu)t}\right)\left[1-\left(\frac{\lambda}{\lambda+\mu}\right)^{x+1}\right]-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}\sum_{k=1}^{x}\sum_{j=1}^{k}\left(\frac{\lambda}{\lambda+\mu}\right)^{k-j}\frac{(\lambda t)^j}{j!},\quad x\geq 0,\ t\geq 0$$

Proof:

First proceeding from Equation 21.12 by induction to get $h(x,t)$ in terms of $h(0,t)$,

$$h(1,t)=\frac{\lambda}{\lambda+\mu}h(0,t)-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}[\lambda t]$$

$$h(2,t)=\frac{\lambda}{\lambda+\mu}h(1,t)-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}\left[\left(\frac{\lambda t}{2!}\right)\right]$$

$$=\frac{\lambda}{\lambda+\mu}\left\{\frac{\lambda}{\lambda+\mu}h(0,t)-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}[\lambda t]\right\}-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}\left[\left(\frac{\lambda t}{2!}\right)\right]$$

$$=\frac{\lambda}{\lambda+\mu}\left\{\left(\frac{\lambda}{\lambda+\mu}\right)^2 h(0,t)-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}\left[\frac{(\lambda t)^2}{2!}\right]\right\}$$

$$h(3,t)=\frac{\lambda}{\lambda+\mu}h(2,t)-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}\left[\frac{(\lambda t)^3}{2!}\right]$$

$$=\frac{\lambda}{\lambda+\mu}\left\{\left(\frac{\lambda}{\lambda+\mu}\right)^2 h(0,t)-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}\left[\frac{(\lambda t)^2}{2!}\right]\right\}-\frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t}\left[\frac{(\lambda t)^3}{3!}\right]$$

$$=\left(\frac{\lambda}{\lambda+\mu}\right)^3 h(0,t)-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}\sum_{j=1}^{3}\left(\frac{\lambda}{\lambda+\mu}\right)^{x-3}\frac{(\lambda t)^3}{j!}$$

…

$$h(k,t)=\left(\frac{\lambda}{\lambda+\mu}\right)^k h(0,t)-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}\sum_{j=1}^{k}\left(\frac{\lambda}{\lambda+\mu}\right)^{k-j}\frac{(\lambda t)^j}{j!},\quad k=1,2,\ldots;t\geq 0 \qquad (21.20)$$

Now to compute the cumulative distribution using Equation 21.20

$$H(x,t)=h(0,t)\sum_{j=0}^{x}\left(\frac{\lambda}{\lambda+\mu}\right)^j-\left(\frac{\mu}{\lambda+\mu}\right)e^{-(\lambda+\mu)t}\sum_{k=1}^{x}\sum_{j=1}^{k}\left(\frac{\lambda}{\lambda+\mu}\right)^{k-j}\frac{(\lambda t)^j}{j!}$$

Consider the first term,

$$h(0,t)\sum_{j=0}^{x}\left(\frac{\lambda}{\lambda+\mu}\right)^j=\frac{\mu}{\lambda+\mu}\left(1-e^{-(\lambda+\mu)}\right)\frac{1-\left(\frac{\lambda}{\lambda+\mu}\right)^{x+1}}{1-\left(\frac{\lambda}{\lambda+\mu}\right)}=\frac{\mu}{\lambda+\mu}\left(1-e^{-(\lambda+\mu)}\right)\left[1-\left(\frac{\lambda}{\lambda+\mu}\right)^{x+1}\right]$$

thus leading to the final result.

22

Planning the Ground Force for Operations in the Post-Cold War Era: A Systems Analysis Approach

W. Peter Cherry
SAIC; Future Combat Systems

Reiner K. Huber
University of the Federal Armed Forces

Thom J. Hodgson
North Carolina State University

22.1 Introduction

In the time since this chapter was originally written in 1999, the political/military environment has grown much more complex. The construct remains valid, and appropriately updated would provide a basis for supporting analysis. During the Cold War period, the force structure of the United States and its NATO allies reflected the requirement to deter and, if necessary, defeat an aggression by the Soviet Union and the Warsaw Pact. It included the capability to deal with "lesser" requirements that had to be addressed occasionally. However, since the end of the Cold War, and Iraq's defeat shortly thereafter, more than 50 military operations have been undertaken to accomplish a series of rather diverse "lesser" missions such as, for example, preventing or ending intrastate conflicts, providing humanitarian assistance and disaster relief, and drug interdiction. This indicates that operations of the "lesser" kind referred to as Stability and Support Operations (SASO) by the US Army have become the rule rather than the exception.

Yet the ground forces of the United States and most of its allies are still configured for Contingency War Operations (CWO) to deal, as a rule, with military threats of the Soviet type, albeit at considerably reduced force levels. The data published by IISS indicate that the manpower level of US ground forces, including marines, has been reduced by 31% since 1989; that of NATO Europe by 23% (Military Balance, 1989/90 and 1996/97). SASO-capabilities are generated, with some delay, as the need arises through reconfiguring and retraining combat arms and support units structured and sized for combined arms combat in CWO. On the other hand, there are a few nations, most notably Canada, which have reconfigured, or are about to do so, their ground forces to deal, as a rule, with SASO demands in a responsive manner, however, depriving them of the capability for participating in CWO lest they mount major efforts for rebuilding their forces.

In fact, today's ground force configurations of the US and Canada may be regarded as representing endpoints of a spectrum of configuration categories that force planners may want to analyze in order to find configurations capable of coping efficiently with the frequent and diverse SASO demands expected in the future, while retaining a sufficient capability for the timely organization of large scale CWO that may be of little likelihood now and in the near future, but may again become a dominant requirement in the long term. In order to assist such an analysis, an analytical approach is presented in this chapter that accounts for the stochastic nature of SASO as revealed by empirical data. From the viewpoint of planning methodology, the fundamental difference between CWO and SASO is that the former imply large-scale deterministic one-time events while latter are smaller in scale, occur repeatedly at random intervals, and vary with regard to length and forces involved. Its application is illustrated by comparing force requirements associated with four basic force configuration options in a scenario characterized by a fictitious simplified demand profile.

22.2 Empirical SASO Data

One important compendium of SASO event data for the US Army, covering the period from 1989 to 1994, has been compiled by Sherwood (1995). The analysis of the data suggests that situations requiring a SASO response are caused by stochastic (Markov) processes of the Poisson-type. Therefore, their arrival can be described in terms of a negative-exponential density function with constant arrival rate. The function implies that the arrival times of individual events are independent, each from another. In reality, this holds true because there is a large number of locations and developments that could possibly require a SASO response. With regard to international conflict, Cioffi-Revilla (1997) has identified 17 conditions sufficient for war and has shown that the distribution of war onset must be Poisson if all conditions occur at constant rates. Similarly, the data suggest that it is not unreasonable to assume that the duration of SASO can be approximated by a negative-exponential density function as well, if events involving "permanent" deployments (e.g. of the Multinational Force and Observers [MNFO]-Sinai) are disregarded. In the context of force planning, SASO events can be considered as permanent if their duration reaches beyond the planning horizon of, say, 2–5 years. In that case they can be addressed in the same manner as the deterministic CWO events (such as major regional contingencies) which require a sufficient force to eventually defeat a given threat.

Thus, with λ denoting the arrival rate of SASO events and μ the rate at which the events are processed, the distribution functions for their arrival time, T_A, and duration time, T_S, can be written as follows:

$$P(T_A \leq t) = 1 - e^{-\lambda t}, \quad E[T_A] = 1/\lambda \tag{22.1a}$$

$$P(T_s \leq t) = 1 - e^{-\mu t}, \quad E[T_s] = 1/\mu \tag{22.1b}$$

Figures 22.1 and 22.2 show histograms, derived from the Sherwood data, of the time between SASO events and of their duration. Accordingly, the arrival rate of SASO events in the period 1989–1994 is approximately 5.3 per year, their average duration slightly more than nine months.

22.3 The Analytic Model

Based on the findings derived from the Sherwood data, of Poisson arrivals of SASO events and negative-exponential event duration, the process can be modeled in the form of an $M/M/n$ queuing system having the capability of handling $i = 1, \ldots, n$ customers and turning away all customers arriving when

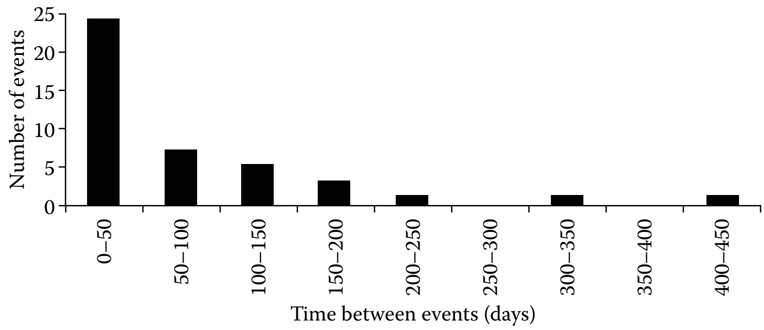


FIGURE 22.1 Histogram of time between arrivals of SASO events.

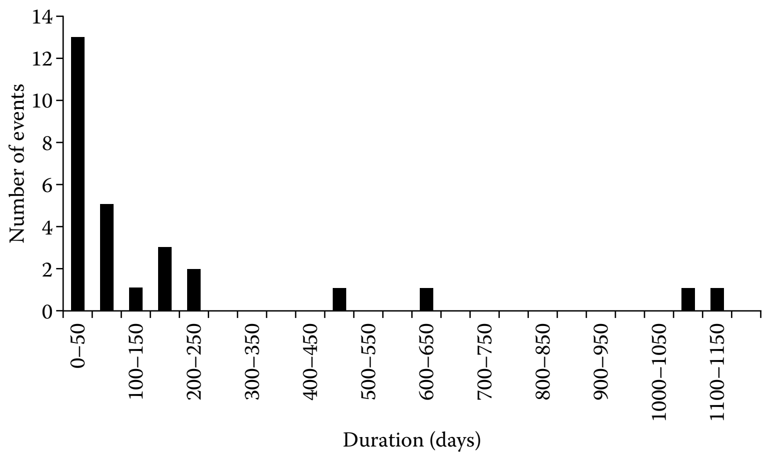


FIGURE 22.2 Histogram of duration of SASO events.

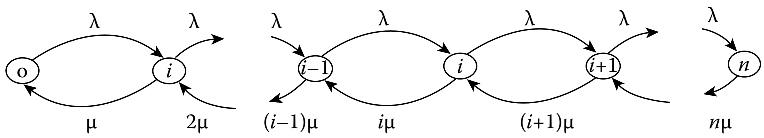


FIGURE 22.3 State transition graph of an *M*/*M*/*n* queuing system.

n customers are being served. In other words, SASO events exceeding the capacity of a SASO force remain unattended.

Figure 22.3 shows a schematic representation of such a system that may be in any one of $i = 1,\ldots,n$ states represented by the nodes. State changes occur either when an event enters the system (at rate λ) or an event is processed (at rate μ) resulting in the following node balance equations:

$$\lambda p(i-1) = i\mu p(i), \quad i = 1, 2, \ldots, n \tag{22.2a}$$

$$\sum_{i=0}^{n} p(i) = 1 \tag{22.2b}$$

The state probabilities $p(i)$ are obtained by successively solving these equations:

$$p(i)=\frac{\frac{a^i}{i!}}{\sum_{k=0}^{n}\frac{a^k}{k!}} \tag{22.3a}$$

with

$$a=\frac{\lambda}{\mu} \tag{22.3b}$$

Thus, if we assume that the SASO force disposes of n units, each capable of handling one SASO event at a time, then the expected fraction of units deployed is given by:

$$\frac{1}{n}\sum_{i=0}^{n} ip(i) \tag{22.4}$$

and the probability that an arriving SASO event is rejected, i.e. not responded to because all n units of the force are deployed, results as:

$$p(n)=\frac{\frac{a^n}{n!}}{\sum_{k=0}^{n}\frac{a^k}{k!}} \tag{22.5}$$

The term "unit" is used in a generic sense. It may refer to any grouping of soldiers or military units required for the respective SASO event. The troop commitment data presented by Sherwood suggest that these requirements are random variables following roughly a geometric distribution. Table 22.1 shows that values resulting from Equations 22.4 and 22.5 for the indicated number of units in a SASO force and the different event arrival rate, given that the units' deployments last a long as the SASO event.

As one would expect, the responsiveness of an SASO force—expressed by the complement of the rejection probability—increases as the force size becomes larger. However, for each arrival rate there is a force size above which comparatively little is to be gained in additional responsiveness. For example, if the marginal increase in responsiveness would have to be at least 5% in order to consider the addition of one further unit to be efficient, then the maximum (efficient) force size in our example would result as $n=4$ for $\lambda=2$, $n=5$ for $\lambda=3$, and $n=7$ for $\lambda=5$.

Another important parameter for repeated SASO is the time between deployments, which must be long enough to permit a unit to recover from the previous and to prepare for the next deployment. Furthermore, the time between deployments in relation to the duration of the deployment reflects upon the balance of troop rotation policies, which should not differ too much for SASO of similar intensity.

The expected time between deployment can be calculated easily enough as a function of the state probabilities. However, when troop rotation policies are superimposed, the calculation becomes considerably more difficult. Therefore, an even-driven Monte Carlo simulation was chosen as a vehicle for exercising the model. Simulation results are compiled in Table 22.2 for 30 cases (10 force sizes and three arrival rates), each for a rotation policy that limits troop deployment to 139 days. In each case, the model was run for 5000 simulated years.

TABLE 22.1 Average Number and Fraction of Units Deployed and Rejection Probability

Arrival Rate	$\lambda = 2$			$\lambda = 3$			$\lambda = 5$		
No. of Units	Average No. Deployed	Average No. Fraction Deployed	Rejection Probability	Average No. Deployed	Average Draction Deployed	Rejection Probability	Average No. Deployed	Average Fraction Deployed	Rejection Probability
1	0.62	0.62	0.62	0.71	0.71	0.71	0.80	0.80	0.80
2	1.09	0.54	0.34	1.31	0.66	0.47	1.55	0.77	0.62
3	1.39	0.46	0.16	1.78	0.60	0.28	2.22	0.74	0.46
4	1.54	0.39	0.06	2.10	0.53	0.15	2.79	0.70	0.32
5	1.61	0.32	0.02	2.30	0.46	0.7	3.25	0.65	0.21
6	1.63	0.27	0.01	2.40	0.40	0.03	3.60	0.60	0.13
7	1.64	0.23	0.00	2.44	0.35	0.01	3.83	0.55	0.07
8	1.64	0.21	0.00	2.46	0.31	0.00	3.97	0.50	0.03

No rotation; expected SASO duration: 300 days.

TABLE 22.2 Average Number of Units Deployed, Average Time Between Deployment, and Rejection Probability

Arrival Rate	λ = 2			λ = 3			λ = 5		
No. of Units	Average No. Deployed	Average Time Between Deployment (days)	Rejection Probability	Average No. Deployed	Average Time Between Deployment (days)	Rejection Probability	Average No. Deployed	Average Time Between Deployment (days)	Rejection Probability
4	148	265	0.06	203	149	0.13	275	75	0.32
5	154	328	0.02	218	187	0.06	315	88	0.19
6	151	375	0	228	226	0.02	346	103	0.11
7	156	468	0	232	264	0.01	366	124	0.06
8	156	544	0	236	311	0	377	145	0.02
9	153	621	0	230	374	0	379	173	0.01
10	157	679	0	229	422	0	383	200	0
11	154	767	0	227	227	0	390	225	0
12	153	847	0	239	239	0	388	253	0
13	157	899	0	233	233	0	388	284	0

Rotation policy: 139 days; expected SASO duration: 300 days.

Suppose that the rotation policy requires a time between deployments that is about twice the deployment time of 139 days. In this case, a force size of $n=4$ units would be required for a SASO arrival rate of $\lambda=2$, $n=7$ for $\lambda=3$, and $n=13$ for $\lambda=5$. One can see that rotation policy is an important factor in determining force requirement for SASO. Compared to the results in Table 22.1 without rotation, this is an increase by a factor of about 1.2 for $\lambda=3$ and 1.6 for $\lambda=5$ in the number of units required for assuring a degree of responsiveness of at least 95%. There is no increase for $\lambda=2$ in our example because the four units required for meeting the effectiveness criterion without rotation satisfy the assumed rotation policy as well.

However, for particular SASO tasks that do not necessarily require military skills, time between deployments of military units can be increased, and as a consequence troop requirements decreased, if initially deployed military units are replaced by civilian contractor elements (CAPCOM) as soon as the situation in the deployment region is considered secure. In order to illustrate these effects, a series of simulation runs was made in which the initial military deployment of 30 days is followed by substitution of civilian personnel and equipment under contract until the SASO event ends. The results show that in comparison to the values in Table 22.2, time between deployments increases, on average, by a factor of three while the average number of military units to be deployed decreases by nearly 90%.

22.4 Ground Force Configuration Options: An Illustrative Example

It has been pointed out earlier that the time between deployments of a unit must be long enough to permit recovery from the previous, and prepare for the next, deployment. Both recovery and preparation time depend on a number of factors, most notably the type and the intensity of operations, the hostility of the operational environment, and the training status of the respective units. When task and operational environment to do not change, some refresher training would be sufficient prior to the next deployment. On the other hand, first assignments to largely novel tasks demands more or less extensive retraining of the respective units.

For example, an armored infantry unit trained for CWO typically undergoes several weeks of training in the skills required for a peacekeeping assignment. Similarly, units returning to their regular CWO tasks following SASO assignments may have to undergo considerable refresher training. Thus, in addition to the skills required for the upcoming mission and the expected operational environment, training and retraining times will be determined largely by what the regular or primary task of a unit is and, therefore, by the configuration of the force.

In a somewhat simplifying manner one may distinguish among four principal categories of ground force configurations:

(1) Option A: A homogeneous force will CWO as its primary mission and retraining part of the force for SASO whenever a demand arises.
(2) Option B: A homogenous force with all of its units being routinely reassigned from CWO to SASO and vice versa.
(3) Option C: A mixed force consisting of specialized units dedicated to CWO and SASO each.
(4) Option D: A homogeneous force with SASO as its primary mission converting (part of) the force when a major war threat begins to emerge.

Option A, with CWO as the primary mission, has the advantage that the type and amount of retraining can be tailored to the actual SASO requirements in a given situation. Thus, there is no retraining loss incurred and, consequently, the maximum possible potential for CWO is available at all times. However, depending on the time required for retraining, responsiveness to crisis situations may be low and, therefore, greatly reduce SASO effectiveness and/or increase SASO force requirements because the crisis situation deteriorated while the earmarked forces were undergoing retraining. Option B, of routinely reassigning units from CWO to SASO, and vice versa, regardless of whether or not there is an actual demand, provides for a high degree of responsiveness since a certain fraction of the force would always be ready for SASO. However, there may be a significant retraining loss as a unit might be reassigned to CWO before the situation calls for SASO. In addition, SASO training would have to address the requirements of a wide range of situation. However, there would only be a marginal loss of CWO capability since ample warning of the emergence of large scale threats should be available in most cases.

Option C, of specialized CWO and SASO forces, assures a high degree of responsiveness in crisis situations calling for SASO as well as vis-à-vis emerging large scale military threats. It should be pointed out that land force specialization at the operations level is not a new idea. For example, the US Army and Marine Corps can be considered as specialized ground force branches of the US armed forces.

Option D, with SASO as the primary mission, assures good responsiveness to several simultaneous SASO demands. This is because more or less the entire force would be available for SASO. Training or retraining for CWO takes place only when a definite threat emerges and, therefore, no "retraining loss" is incurred. However, depending on whether or not the units have undergone some initial (basic) CWO training, there would be little or no capability for addressing eventual "pop-up" CWO threats emerging on short notice.

A sizeable analytical effort has yet to be undertaken to develop, on the level of relevant SASO tasks, the data necessary for realistic assessment of configuration options and their embedded rotation policies. The work at George Mason University (1996) may be considered as a first step. Among others, it has compiled a comprehensive list of almost 400 individual SASO tasks and analyzed 29 peace operations to determine their relevance for the principal SASO categories. Thus, in the context of this chapter, we shall make some plausible assumptions about aggregate unit-level SASO training and recovery times relative to the duration of deployments, as depicted in Table 22.3, for three mission environments characterized by the complexity of the SASO task (normal: N, complex: C) and the attitude of the population in the mission area (friendly: F; neutral: N hostile: H). The SASO demand profile is defined in terms of the arrival rate and magnitude of SASO events under different mission environments.

For our illustrative example, we assume a rotation policy based on a deployment time of 139 days, as in Table 22.2, and training and recovery time requirements (i.e. time between deployments) as specified in Table 22.3. The demand profile expects SASO events in an NF environment at an arrive rate of $\lambda=5$, in a CN environment at $\lambda=3$, and in a CH environment at $\lambda=2$. The SASO force requirement resulting with these assumptions for the four principal force configuration options are presented in Table 22.4.

TABLE 22.3 SASO Training Time (T_T) and Recovery Rime (T_R) Requirements Relative to the Deployment Time (T_D)

	Mission Environment					
	N: normal, F: friendly		C: complex, N: neutral		C: complex, H: hostile	
Configuration Option	T_T	T_R	T_T	T_R	T_T	T_R
A	1.5	1.5	3.0	1.6	3.0	3.0
B	1.0	1.5	2.0	1.6	2.0	3.0
C	0.5	0.5	1.0	0.6	1.0	1.0
D	0.5	0.5	1.0	0.6	1.0	1.0

TABLE 22.4 Force Configuration and SASO Force Requirements

	SASO Demand Profile			
Configuration Option	NF ($\lambda=5$)	CN ($\lambda=3$)	CH ($\lambda=2$)	Total
A	19	16	12	47
B	16	12	10	38
C	8	6	4	18
D	8	6	4	18

The results reflect the sensitivity of requirements towards the time between deployment needed for recovery and (re)training. Therefore, option A, of a homogeneous CWO force dealing with SASO events on a case by case basis, turns out to be the most demanding. It requires 47 units for coping with the assumed SASO demand profile as opposed to 18 when specialized units dedicated to SASO are available in a mixed force (option C). Thus, regardless of what total force size above 47 units the budget may allow, a mixed force can be expected not only to be more responsive to SASO demands, but also have a higher number of units available for CWO on short notice without having to abort ongoing SASO. To a lesser degree, this conclusion would also be the case for option B, featuring a homogeneous force being routinely reassigned and retrained, from CWO to SASO and vice versa. In our example, option B unit requirements turn out to be more than twice those of option C, however, about 20% below those of option A.

The total force requirements shown in Table 22.4 would decrease about 30% for options A and B, and by about 20% for options C and D, if the initially deployed military units were replaced by CAPCOM after 30 days in all SASO events arriving in an NF environment, and in two thirds of the SASO events arriving in a CN environment.

22.5 Preliminary Conclusions

There may be strong arguments in favor of maintaining a homogeneous force at this time. However, with a view to the demands for SASO on one hand, and a maximum capability for handling ad hoc CWO threats on the other, our preliminary analysis suggests that the force should be trained and retrained on a routine basis for both CWO and SASO, and civilian contractor elements should replace military SASO forces whenever the mission environment permits. In the long term, a mixed force of specialized units for CWO and SASO appears to be the preferable ground force configuration, in particular since its ad hoc CWO capability seems to be less sensitive to manpower reductions and budget cuts.

To a large extent, the "specialized" units could consist of civilian personnel and equipment deployed under contract only when needed. Option D may be considered as a possible ground force configuration for a time when the threat of major war emerging on short notice may have disappeared altogether. This also applies to wars similar to the war in Bosnia. There is no doubt that a necessary condition for its prevention would have been a credible deterrent in form of a sufficient and highly visible CWO capability of the international community (see Huber, 1994, and Huber and Miller, 1996). Then its viability will depend largely on whether a CWO capability can be reconstituted within the effective warning time should major war threats emerge again.

References

Cioffi-Revilla, C. 1997. The Road to War: Modeling International Conflict with Principles of Political Uncertainty Theory. Proceedings of Conference on Advances in Modelling of International Conflict, Wadham College, Oxford, 13–15 April.

Cohen, A.C. 1965. Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples. *Technometrics,* 7(4).

George Mason University. 1996. Analysis of Peace Operations Tasks: Relevance and Factor Analysis. Research Report, Fairfax, The Institute of Public Policy.

Harter, H.L., and A.H. Moore. 1965. Maximum-likelihood estimation of the parameters of gamma and Weibull populations from complete and from censored samples. *Technometrics,* 7(4).

Hartley III, D.S. 1996. Operations Other than War: Requirements for Analysis Tools. Report K/DSRD-1098, Lockheed Martin Energy Systems, Inc.

Huber, R.K. 1994. Deterrence in Bosnia-Herzegovina: A retrospective analysis of military requirements before the war. *European Security*, 3(3), 541–551.

Huber, R.K., and G. Miller. 1996. Strategies for peace support operations: A decision-theoretic analysis. In: A.E.R. Wooddock, and D.F. Davis, Eds., *Analytic Approaches to the Study of Future Conflict*. Halifax: The Canadian Peacekeeping Press, 227–237.

National Guard Bureau Operations Center. State Active Duty Report, 1994–1995.

Requirements Validation Database. ODCSLOG, DA, 13 June 1995.

Saaty, T.L. 1961. *Elements of Queuing Theory with Application*. New York: McGraw-Hall.

Sherwood, J.D. 1995. US Army Operations Other than War Since 1989. US Army Center of Military History, April.

V

Supply Chain and Decision Making

23

Supply Chain Management*

Douglas M. Lambert
The Ohio State University

23.1 Overview

There is a great deal of confusion regarding what supply chain management involves [1]. In fact, many people using the name supply chain management treat it as a synonym for logistics or as logistics that includes customers and suppliers [2]. Others view supply chain management as the new name for purchasing or operations [3], or the combination of purchasing, operations and logistics. However, successful supply chain management requires cross-functional integration within the firm and across the

* This chapter is from Douglas M. Lambert, Editor. *Supply Chain Management: Processes, Partnerships, Performance*, 3rd Edition. Sarasota, FL: Supply Chain Management Institute, 2008. Copyright 2008, Supply Chain Management Institute, used with permission. For more information, see: http://www.scm-institute.org.

network of firms that comprise the supply chain. The challenge is to determine how to successfully accomplish this integration.

In this chapter, supply chain management (SCM) is defined and the uniqueness of our framework is explained. Descriptions of the transactional and the relationship management views of business process management are provided. The SCM processes are described as well as the importance of standard business processes. Then, there is an explanation of how the SCM processes can be used to achieve cross-functional and cross-firm integration. There is a description of how customer relationship management and supplier relationship management form the critical SCM linkages and how their impact on the financial performance of the organization can be measured. Also, the partnership model is introduced as means of building high-performance relationships in the supply chain. The ultimate success of the business will depend on management's ability to integrate the company's intricate network of business relationships.

23.2 Introduction

One of the most significant paradigm shifts of modern business management is that individual businesses no longer compete as solely autonomous entities, but rather within supply chains. In this emerging competitive environment, the ultimate success of the business will depend on management's ability to integrate the company's intricate network of business relationships [5]. Increasingly the management of relationships across the supply chain is being referred to as SCM. Strictly speaking, the supply chain is not a chain of businesses, but a network of businesses and relationships. SCM offers the opportunity to capture the synergy of intra and intercompany integration and management. In that sense, SCM deals with business process excellence and represents a new way of managing the business and relationships with other members of the supply chain.

Thus far, there has been relatively little guidance from academia, which has in general been following rather than leading business practice [6]. There is a need for building theory and developing normative tools and methods for successful SCM practice. The Global Supply Chain Forum (GSCF), a group of noncompeting firms and a team of academic researchers, has been meeting regularly since 1992 with the objective to improve the theory and practice of SCM. The definition of SCM developed and used by the members of GSCF follows [7]:

> Supply chain management is the integration of key business processes from end-user through original suppliers that provides products, services, and information that add value for customers and other stakeholders.

This view of SCM is illustrated in Figure 23.1, which depicts a simplified supply chain network structure, the information and product flows, and the SCM processes that integrate functions within the company as well as other firms across the supply chain. Thus, standard SCM processes are necessary to manage the links across intra and intercompany boundaries.

This chapter is organized as follows. First there is a description of what SCM is not. This is followed by a section describing SCM and a brief overview of business process management. Next, the SCM processes are described. Then, the need for standard business processes is introduced. Next, the need for cross-functional and cross-firm involvement in the SCM processes is described. This is followed by a section that illustrates how customer relationship management and supplier relationship management form the critical SCM linkages. Then, you will be shown how to measure the financial impact of customer relationship management and supplier relationship management. Building high-performance relationships in the supply chain and a summary of the SCM framework are the last two topics covered. Finally, conclusions are presented.

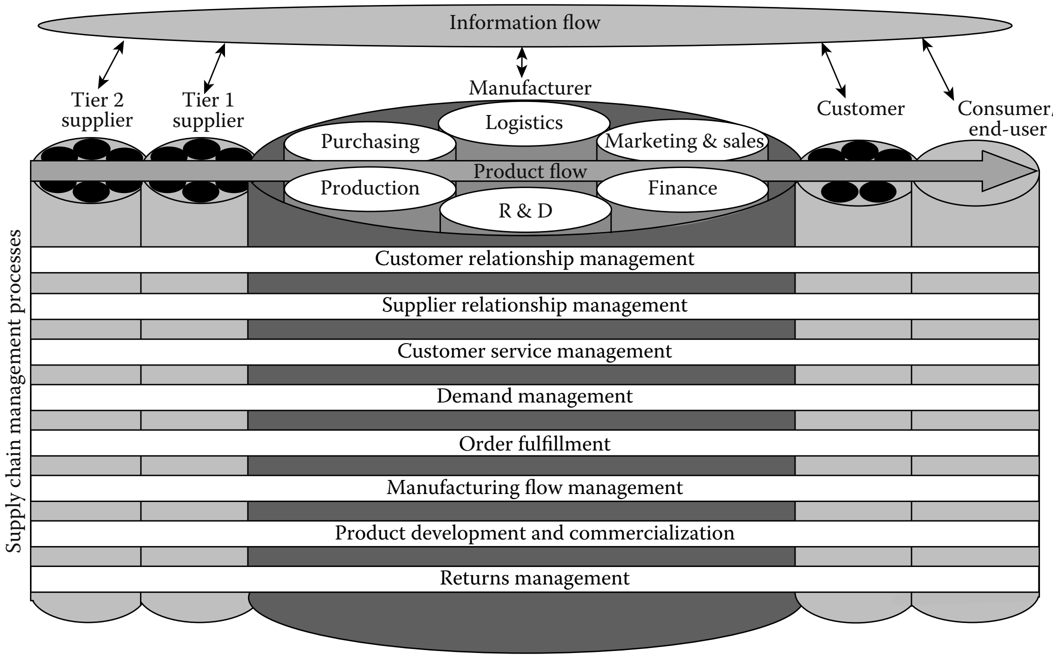


FIGURE 23.1 Supply chain management: Integrating and managing business processes across the supply chain. (Adapted from Douglas M. Lambert, Martha C. Cooper, and Janus D. Pagh. 1998. Supply chain management: Implementation issues and research opportunities. *The International Journal of Logistics Management*, 9(2), 2.)

23.3 What SCM is Not

The term SCM was originally introduced by consultants in the early 1980s [8] and subsequently has become widely used [9]. Since the late 1980s, academics have attempted to give structure to SCM [10]. Bechtel and Jayaram [11] provided an extensive retrospective review of the literature and research on SCM. They identified generic schools of thought, and the major contributions and fundamental assumptions of SCM that must be challenged in the future. Until recently most practitioners [12], consultants [13] and academics [14] viewed SCM as not appreciably different from the contemporary understanding of logistics management, as defined by the Council of Logistics Management (CLM) in 1986 [15]. That is, SCM was viewed as logistics that was integrated with customers and suppliers. Logistics as defined by the CLM always represented a supply chain orientation, "from point-of-origin to point-of-consumption." Then, why the confusion? It is probably due to the fact that logistics is a function within companies and is also a bigger concept that deals with the management of material and information flows across the supply chain. This is similar to the confusion over marketing as a concept and marketing as a functional area. Thus, the quote from the CEO who said, "Marketing is too important to be left to the marketing department." Everybody in the company should have a customer focus. The marketing concept does not apply just to the marketing department. Everyone in the organization should focus on serving the customer's needs at a profit.

The understanding of SCM has been reconceptualized from integrating logistics across the supply chain to integrating and managing key business processes across the supply chain. Based on this emerging distinction between SCM and logistics, in 2003, CLM announced a modified definition of logistics. The modified definition explicitly declares CLM's position that logistics management is only a part of SCM. The revised definition follows:

> Logistics is that part of SCM that plans, implements, and controls the efficient, effective forward and reverse flow and storage of goods, services, and related information between the point-of-origin and the point-of-consumption in order to meet customers' requirements [16].

SCM is not just confused with logistics. Those in the operations management area, such as APICS, are renaming what they do as SCM as are those working in the procurement area [17]. Some universities have created departments of SCM by combining purchasing, operations and logistics faculty, which is a perspective that has appeared in the literature [18].

23.4 Just What is Supply Chain Management?

Figure 23.1 shows the supply chain network structure of a manufacturer with two tiers of customers and two tiers of suppliers, the information and product flows, and the SCM processes that must be implemented within organizations across the supply chain. All of the processes are cross-functional and cross-firm in nature. Every organization in the supply chain needs to be involved in the implementation of the same eight processes but corporate silos and functional silos within companies are barriers to this integration (see Figure 23.1). In most major corporations, functional managers are rewarded for behavior that is not customer friendly or shareholder friendly. This is because the metrics used focus on functional performance such as cost per case, asset utilization, and revenue goals, not on customer value or shareholder value. Successful management of the supply chain requires the involvement of all of the corporate business functions [19]. A network of companies cannot be managed with fewer functions than are necessary to manage one company.

In his keynote address to the International Association of Food Industry Suppliers in March 2005, Tom Blackstock, Vice President of Supply Chain Operations at Coca-Cola North America, confirmed the need to involve all business functions in supply chain management when he said: "Supply chain management is everybody's job" [20]. In 2006, John Gattorna expressed a similar perspective on the breadth of management necessary for successful implementation of SCM:

> We have to embrace a far more liberal view of the supply chain. In effect, the supply chain is any combination of processes, functions, activities, relationships, and pathways along which products, services, information, and financial transactions move in and between enterprises. It also involves any and all movement of these from original producer to ultimate end-user or consumer, and everyone in the enterprise is involved in making this happen [21].

In reality, a supply chain is much more complex than the row of silos depicted in Figure 23.1. For a company in the middle of the supply chain like a consumer goods manufacturer, the supply chain looks like an uprooted tree (see Figure 23.2) where the root system represents the supplier network and the branches of the tree represent the customer network. The supply chain will look different depending on a firm's position in it. For example, in the case of a retailer, like Wal-Mart, the consumers would be next to the dark square (Wal-Mart) in the center of Figure 23.2 making them the only tier in the customer network. For an initial supplier, such as a shrimper, there would be no suppliers associated with the product flow.

Managing the entire supply chain is a very challenging task. Managing all suppliers back to the point-of-origin and all products/services out to the point-of-consumption might appear to be overwhelming. It is probably easier to understand why executives would want to manage their supply chains to the point-of-consumption because whoever has the relationship with the end-user has the power in the supply chain. Intel created a relationship with the end-user by having computer manufacturers place an "Intel inside" label on their computers. This affects the computer manufacturer's ability to switch microprocessor suppliers. However, opportunities exist to significantly improve profits by managing the supplier network as well. For example, Coca-Cola is one of the largest purchasers

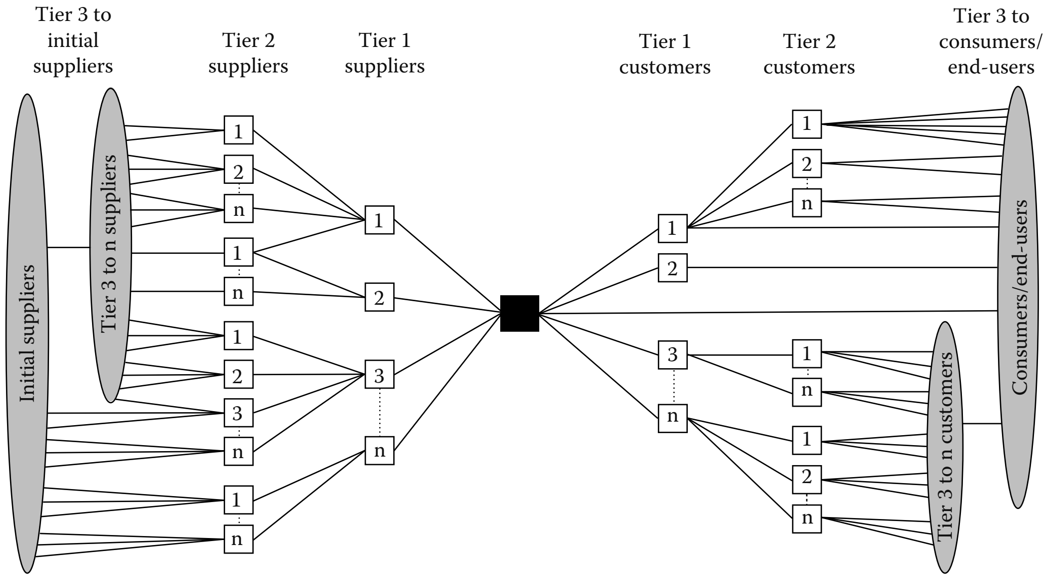


FIGURE 23.2 Supply chain network structure. (Adapted from Douglas M. Lambert, Martha C. Cooper, and Janus D. Pagh. 1998. Supply chain management: Implementation issues and research opportunities. *The International Journal of Logistics Management*, 9(2), 3.)

of PET resins in the world as a result of managing its suppliers of packaging materials beyond Tier 1. Because resin costs represent such a large portion of the package cost, Coca-Cola contracts for PET resins directly with the resin producer. This practice also results in improved availability and less price volatility.

At the end of the day, SCM is about relationship management. A supply chain is managed, link-by-link, relationship-by-relationship, and the organizations that manage these relationships best will win. The links in the chain are formed by the customer relationship management process of the seller organization and the supplier relationship management process of the buyer organization. The focus of the remainder of this chapter is on business process management, the eight SCM processes, how customer relationship management and supplier relationship management form the linkages for integrating companies in the supply chain, how the financial impact of each linkage is measured, and how a tool known as the partnership model, can be used to structure relationships with key customers and suppliers.

23.5 Business Process Management

Increasingly, managers want to implement business processes and integrate them with other key members of the supply chain [22]. A business process is a structured set of activities with specified business outcomes for customers [23]. Initially, business processes were viewed as a means to integrate corporate functions within the firm. Now, business processes are used to structure the activities between members of a supply chain. Hammer has pointed out that it is in the integration of business processes across firms in the supply chain where the real "gold" can be found [24].

The concept of organizing the activities of a firm as business processes was introduced in the late 1980s [25] and became popular in the early 1990s, after the publication of books by Hammer and Champy [26], and Davenport [27]. The motivation for implementing business processes within and across members of the supply chain might be to make transactions efficient and effective, or to structure inter-firm relationships in the supply chain. Both approaches are customer oriented. The first focuses on meeting the customer's expectations for each transaction, the other on achieving longer term mutual fulfillment of promises.

The transactional view of business process management is rooted in advances in information and communication technology which enabled time compression and availability of information throughout the organization [28]. The focus is not on automating the established business processes, but on redesigning businesses [29]. This transactional approach to redesigning business processes is based on Taylor's principles of scientific management which aim to increase organizational efficiency and effectiveness using engineering principles from manufacturing operations [30]. In this case, business process redesign is based on standardizing transactions and the transfer of information [31]. The goal is to improve outcomes for customers by making transactions more efficient and accurate.

The second view of business process management focuses on managing relationships in the supply chain and is based on an evolving view from the field of marketing. A significant amount of the marketing literature is concerned with market transactions (business transactions with customers) and the fulfillment of orders. Rooted in economic theory, researchers studied the efficiency of transactions with the customers, which raised awareness about the importance of customer retention [32]. Obtaining repeat business, that is to conduct multiple transactions with the same customer, is more cost efficient than obtaining a new customer [33]. This view of marketing, managing transactions with customers, is dominated by the "4Ps": product, price, promotion, and place [34].

The early marketing channels researchers such as Alderson and Bucklin conceptualized why and how channels are created and structured [35]. From a supply chain standpoint, these researchers were on the right track in terms of: (1) identifying who should be a member of the marketing channel; (2) describing the need for channel coordination; and (3) drawing actual marketing channels. However, for the last 40 years most marketing channels researchers ignored two critical issues. First, they did not build on the early contributions by including suppliers to the manufacturer, and thus neglected the importance of a total supply chain perspective. Unlike the marketing channels literature, a major weakness of much of the SCM literature is that the authors appear to assume that everyone knows who is a member of the supply chain. There has been little effort to identify specific supply chain members, key processes that require integration or what management must do to successfully manage the supply chain. Second, channels researchers focused on marketing activities and flows across the marketing channel, and overlooked the need to integrate and manage cross-functionally within and across companies. In 1992, Webster [36] challenged marketers and marketing researchers to consider relationships with multiple firms. He also called for cross-functional consideration in strategy formulation.

During the 1990s, a paradigm shift occurred with the introduction of the concept of relationship marketing [37]. The goal of relationship marketing "…is to establish, maintain, and enhance… relationships with customers and other partners, at a profit, so that the objectives of the parties involved are met. This is achieved by mutual exchange and fulfillment of promises" [38]. Thus, the focus of developing and maintaining relationships in the supply chain is beyond the fulfillment of one or a set of transactions. In the new environment, managers need to focus on helping customers achieve their objectives.

The field of relationship marketing is focused on the customer-side, looking downstream in the supply chain. However, the development and maintenance of relationships with key suppliers should be based on the same pillars, mutuality and fulfillment of promises, in order for suppliers to be profitable. Management needs the support of the firm's key suppliers to fulfill the promises made to customers and meet financial goals. In other words, corporate success is based on relationship management with both suppliers and customers. The management of interorganizational relationships with members of the supply chain involves people, organizations, and processes [39]. In fact, the ability to manage interorganizational relationships "… may define the core competence of some organizations as links between their vendors and customers in the value chain" [40].

In 1992, executives from a group of international companies and a team of academic researchers, began development of a relationship oriented and process-based SCM framework. In February 1996, the The Global Supply Chain Forum (GSCF) framework was presented in a three-day executive seminar co-sponsored by the CLM, and was later presented in the literature [41]. The eight GSCF processes are cross-functional and are meant to be implemented interorganizationally across key members of the supply chain. The motivation for developing the framework was to provide structure to assist academics with their research on supply chain management and practitioners with implementation.

23.6 The Supply Chain Management Processes

Empirical research has led to the conclusion that "the structure of activities within and between companies is a critical cornerstone of creating unique and superior supply chain performance" [42]. In our research, executives believed that competitiveness and profitability could increase if key internal activities and business processes are linked and managed across multiple companies. Thus, "corporate success requires a change from managing individual functions to integrating activities into SCM processes" [43]. In many major corporations, such as Coca-Cola, management has reached the conclusion that optimizing the product flows cannot be accomplished without implementing a process approach to the business [44]. Several authors have suggested implementing business processes in the context of SCM, but there is not yet an "industry standard" on what these processes should be. The value of having standard business processes in place is that managers from organizations across the supply chain can use a common language and can link-up their firms' processes with other members of the supply chain, as appropriate. The SCM processes identified by The GSCF and shown in Figure 23.1 are:

- Customer relationship management
- Supplier relationship management
- Customer service management
- Demand management
- Order fulfillment
- Manufacturing flow management
- Product development and commercialization
- Returns management

Each SCM process has both strategic and operational subprocesses. The strategic subprocesses provide the structure for how the process will be implemented and the operational subprocesses provide the detailed steps for implementation. The strategic process is a necessary step in integrating the firm with other members of the supply chain, and it is at the operational level that the day-to-day activities take place. Each process is led by a management team that is comprised of managers from each business function, including: marketing, sales, finance, production, purchasing, logistics and, research and development. Teams are responsible for developing the procedures at the strategic level and for implementing them at the operational level. A brief description of each of the eight processes follows.

23.6.1 Customer Relationship Management

The customer relationship management process provides the structure for how the relationships with customers will be developed and maintained. Management identifies key customers and customer groups to be targeted as part of the firm's business mission. These decisions are made by the leadership team of the enterprise and the owner of the strategic process is the CEO. The goal is to segment customers based on their value over time and increase customer loyalty of target customers by providing customized products and services. Cross-functional customer teams tailor Product and Service Agreements (PSAs) to meet the needs of key accounts and for segments of other customers. The PSAs specify levels of performance. The teams work with key customers to improve processes and reduce demand variability and nonvalue-added activities. Performance reports are designed to measure the profitability of individual customers as well as the firm's impact on the financial performance of the customer [45].

23.6.2 Supplier Relationship Management

The supplier relationship management process provides the structure for how relationships with suppliers will be developed and maintained. As the name suggests, this is a mirror image of customer relationship management. Just as a company needs to develop relationships with its customers, it also needs to foster relationships with its suppliers. Close relationships are developed with a small subset of suppliers based on the value that they provide to the organization over time, and more traditional relationships are maintained with the others. Supplier teams negotiate PSAs with each key supplier that defines the terms of the relationship. For each segment of less critical suppliers, a standard PSA is provided and it is not negotiable. Supplier relationship management is about defining and managing these PSAs. Partnerships are developed with a small core group of suppliers. The desired outcome is a win-win relationship where both parties benefit.

23.6.3 Customer Service Management

Customer service management is the SCM process that deals with the administration of the PSAs developed by customer teams as part of the customer relationship management process. Customer service managers monitor the PSAs and proactively intervene on the customer's behalf if there is going to be a problem delivering on promises that have been made. The goal is to solve problems before they affect the customer. Customer service managers will interface with other process teams, such as supplier relationship management and manufacturing flow management to ensure that promises made in the PSA's are delivered as planned.

23.6.4 Demand Management

Demand management is the SCM process that balances the customers' requirements with the capabilities of the supply chain. With the right process in place, management can match supply with demand proactively and execute the plan with minimal disruptions. The process is not limited to forecasting. It includes synchronizing supply and demand, reducing variability and increasing flexibility. For example, it involves managing all of the organization's practices that increase demand variability, such as end-of-quarter loading and terms of sale which encourage volume buys. A good demand management process uses point-of-sale and key customer data to reduce uncertainty and provide efficient flows throughout the supply chain. Marketing requirements and production plans should be coordinated on an enterprise-wide basis. In advanced applications, customer demand and production rates are synchronized to manage inventories globally.

23.6.5 Order Fulfillment

The order fulfillment process involves more than just filling orders. It includes all activities necessary to design a network and enable a firm to meet customer requests while minimizing the total delivered cost. At the strategic level, for example, it is necessary to determine which countries should be used to service the needs of various customers considering service requirements, tax rates and where profits should be earned as well as import and export regulations. While much of the actual work will be performed by the logistics function, it needs to be implemented cross-functionally and with the coordination of key suppliers and customers. The objective is to develop a seamless process from the various customer segments to the organization and then on to its suppliers.

23.6.6 Manufacturing Flow Management

Manufacturing flow management is the SCM process that includes all activities necessary to obtain, implement and manage manufacturing flexibility in the supply chain and to move products into, through and out of the plants. Manufacturing flexibility reflects the ability to make a wide variety of products in a timely manner at the lowest possible cost. To achieve the desired level of manufacturing flexibility, planning and execution must extend beyond the four walls of the manufacturer to other members of the supply chain.

23.6.7 Product Development and Commercialization

Product development and commercialization is the SCM process that provides the structure for developing and bringing to market products jointly with customers and suppliers. Effective implementation of the process not only enables management to coordinate the efficient flow of new products across the supply chain, but also assists other members of the supply chain with the ramp-up of manufacturing, logistics, marketing and other activities necessary to support the commercialization of the product. The product development and commercialization process team must coordinate with customer relationship management process teams to identify customer articulated and unarticulated needs; select materials and suppliers in conjunction with the supplier relationship management process teams; and, work with the manufacturing flow management process team to develop production technology to manufacture and implement the best product flow for the product/market combination.

23.6.8 Returns Management

Returns management is the SCM process by which activities associated with returns, reverse logistics, gate-keeping, and avoidance are managed within the firm and across key members of the supply chain. The correct implementation of this process enables management not only to manage the reverse product flow efficiently, but to identify opportunities to reduce unwanted returns and to control reusable assets such as containers. While significant opportunities to reduce costs are possible through better management of reverse logistics, even greater potential to reduce costs and increase revenue are possible by eliminating those management practices and performance failures that cause returns.

23.7 The Requirement for Standard Business Processes

Thousands of activities are performed and coordinated within a company, and every company is by nature in some way involved in supply chain relationships with other companies [46]. When two companies build a relationship, some of their internal activities will be managed between the two companies [47]. Since both companies have linked some internal activities with other members of their supply chain, a link between two companies is thus a link in what might be conceived as a supply chain network. For example, the internal activities of a manufacturer can affect the internal activities of a distributor, which in turn have an effect on the internal activities of a retailer. Ultimately, the internal activities of the retailer are linked with and can affect the activities of the end-user.

Our research team has found that in some companies, executives emphasize a functional structure (see Figure 23.3, Tier 1 Supplier) and others a process structure (see Figure 23.3, Manufacturer, Tier 2 Supplier and both tiers of customers). Those companies with processes had different numbers of processes consisting of different activities and links between activities. Different names were used for similar processes, and similar names for different processes. This lack of inter-company consistency is a cause of significant friction and inefficiencies in supply chains. It is important that managers in different firms speak the same language (use the same terminology). There is generally an understanding of what

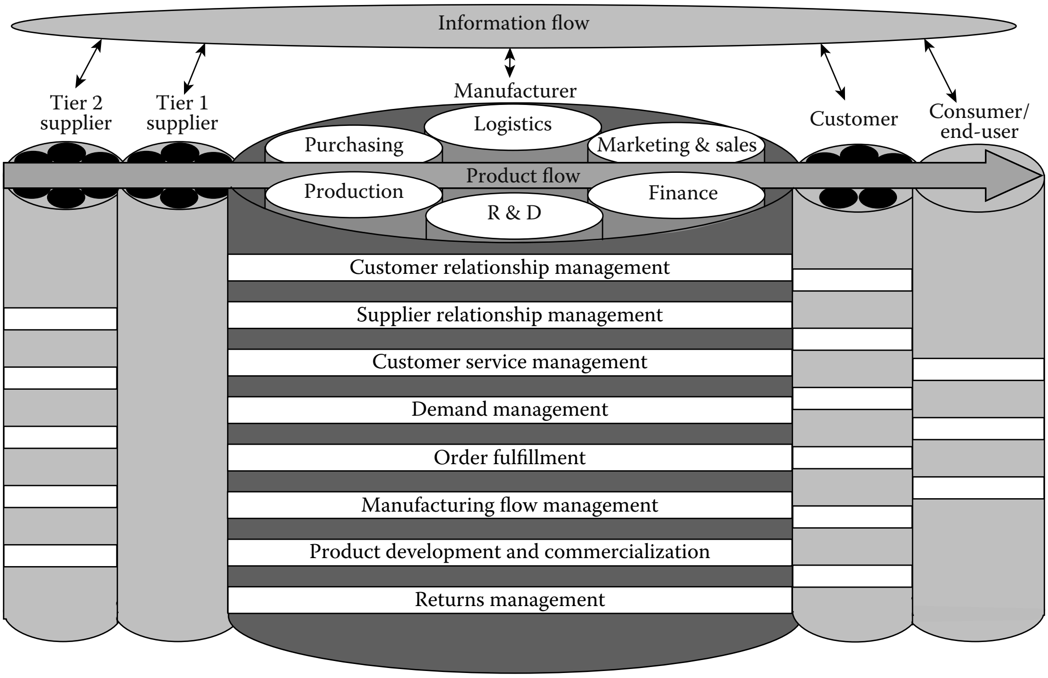


FIGURE 23.3 Supply chain management: The disconnects. (Adapted from Douglas M. Lambert, Martha C. Cooper, and Janus D. Pagh. 1998. Supply chain management: Implementation issues and research opportunities. *The International Journal of Logistics Management*, 9(2), 10.)

corporate functions like marketing, manufacturing and finance represent. If management in each firm identifies its own set of processes, how can these processes be linked across firms?

In some of the GSCF companies, business processes extended to suppliers and were managed to some extent between the two firms involved. This may imply that when a leadership role is taken, firms in the supply chain will use the same business processes. When this is possible, each member of the band is playing the same tune.

The number of business processes that should be integrated and managed between companies will likely vary. However, in each specific case, it is important that executives thoroughly analyze and discuss which key business processes to integrate and manage.

23.8 Achieving Cross-functional and Cross-firm Involvement Using the Supply Chain Management Processes

If the proper coordination mechanisms are not in place across the various functions, the process will be neither effective nor efficient. By taking a process focus, all functions that touch the product or are involved in the service delivery must work together. Figure 23.4 shows examples of how managers from each function within the organization provide input to the eight SCM processes. For example, in the customer relationship management process, marketing provides the knowledge of customers and marketing programs as well as the budget for marketing expenditures, sales provides the account management expertise, research and development provides the technological capabilities to develop product solutions that meet customer requirements, logistics provides knowledge of logistics and customer service capabilities, production provides the manufacturing capabilities, purchasing provides knowledge of supplier capabilities, and finance provides customer profitability reports. Customers and suppliers

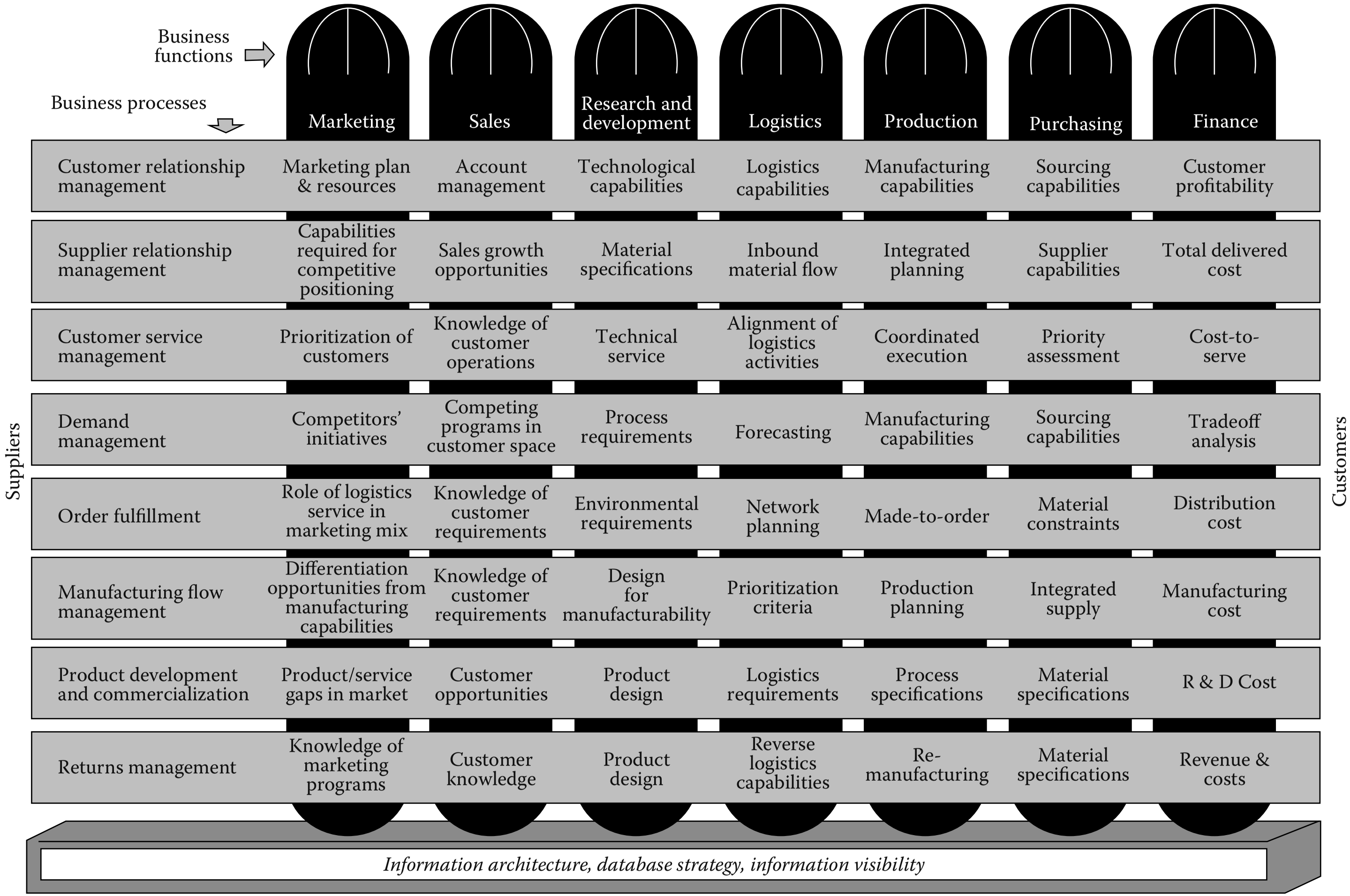


FIGURE 23.4 Functional involvement in the supply chain management processes. Note: Process sponsorship and ownership must be established to drive the attainment of the supply chain vision and eliminate the functional silo mentality. (Adapted from Keely L. Croxton, Sebastian J. García-Dastuque, and Douglas M. Lambert. 2001. The supply chain management processes. *The International Journal of Logistics Management*, 12(2), 31.)

are shown in Figure 23.4 to make the point that each of these processes, to be properly implemented, requires the involvement of all business functions as well as customers and suppliers. When third-party logistics providers are used, representatives from these firms should serve on the process teams to provide their logistics expertise.

If the proper coordination mechanisms are not in place across the various functions, the process will be neither effective nor efficient. By taking a process focus, all functions that touch the product or are involved in the service delivery must work together. Figure 23.4 shows examples of how managers from each function within the organization provide input to the eight SCM processes. For example, in the customer relationship management process, marketing provides the knowledge of customers and marketing programs as well as the budget for marketing expenditures, sales provides the account management expertise, research and development provides the technological capabilities to develop product solutions that meet customer requirements, logistics provides knowledge of logistics and customer service capabilities, production provides the manufacturing capabilities, purchasing provides knowledge of supplier capabilities, and finance provides customer profitability reports. Customers and suppliers are shown in Figure 23.4 to make the point that each of these processes, to be properly implemented, requires the involvement of all business functions as well as customers and suppliers. When third-party logistics providers are used, representatives from these firms should serve on the process teams to provide their logistics expertise.

In order to achieve cross-firm integration, management needs to choose the type of relationship that is appropriate for each link in the supply chain [48]. Not all links throughout the supply chain should be closely coordinated and integrated. The most appropriate relationship is the one that best fits the specific set of circumstances [49].

Determining which members of the supply chain deserve management attention is based on their importance to the firm's success. In some companies, management works closely with second-tier members of the supply chain in order to achieve specific supply chain objectives, such as product availability, improved quality, improved product introductions, or reduced overall supply chain costs. For example, a tomato ketchup manufacturer in New Zealand conducts research on tomatoes in order to develop plants that provide larger tomatoes with fewer seeds. Their contracted growers are provided with young plants in order to ensure the quality of the output. Since the growers tend to be small, the manufacturer negotiates contracts with suppliers of equipment and agricultural chemicals such as fertilizer and pesticides. The farmers are encouraged to purchase materials and machinery using the manufacturer's contract rates. This results in higher quality tomatoes and lower prices without sacrificing the margins and financial strength of the growers.

23.9 The Critical Supply Chain Management Linkages

Customer relationship management and supplier relationship management form the critical linkages throughout the supply chain (see Figure 23.5). For each supplier in the supply chain, the ultimate measure of success for the customer relationship management process is the positive change in profitability of an individual customer or segment of customers over time. For each customer, the most comprehensive measure of success for the supplier relationship management process is the impact that a supplier or supplier segment has on the firm's profitability. The goal is to increase the profitability of each organization by developing the relationship. The biggest potential roadblock is failure to reach agreement on how to split the gains that are made through joint improvement efforts. The overall performance of the supply chain is determined by the combined improvement in profitability of all of its members from one year to the next.

Typically, large sums of money are spent by corporations to attract new customers; yet these same companies are often complacent when it comes to nurturing existing customers to build and strengthen relationships with them [50]. However, for most companies existing customers represent the best opportunities for profitable growth. There are direct and strong relationships between profit growth; customer loyalty; customer satisfaction; and, the value of goods delivered to customers [51]. In a business-to-business environment, the customer relationship management process is the supply chain management

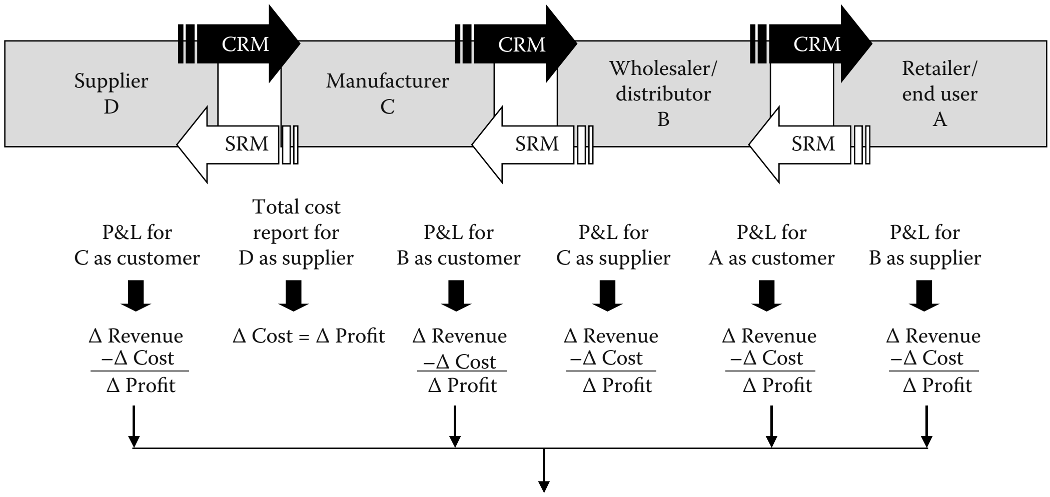


FIGURE 23.5 Customer relationship management (CRM) and supplier relationship management (SRM): The critical supply chain management linkages. (Adapted from Douglas M. Lambert and Terrance L. Pohlen. 2001. Supply chain metrics. *The International Journal of Logistics Management*, 12(1), 14.)

process that provides the structure for how relationships with customers are developed and maintained. The decision regarding who is a key customer requires evaluation of the profitability and potential profitability of individual customers. Then, cross-functional customer teams tailor PSAs to meet the needs of key accounts and segments of other customers [52]. PSAs come in many forms, both formal and informal, and may be referred to by different names. However, for best results they should be formalized as written documents. Teams work with key accounts to improve processes, and reduce demand variability and nonvalue-added activities.

Supplier relationship management is the mirror image of customer relationship management. All suppliers are not the same. Some suppliers contribute disproportionately to the firm's success and with these organizations, it is important to have cross-functional teams interacting. There will be teams established for each key supplier and for each segment of nonkey suppliers. The teams are comprised of managers from several functions, including marketing, finance, research and development, production, purchasing and logistics. At the strategic level, the team is responsible for developing the strategic process, and seeing that it is implemented. Supplier teams have the day-to-day responsibility for managing the process at the operational level.

Since, there will be a customer team and a supplier team for each key customer and supplier, and for each segment of nonkey customers/suppliers, it is important that teams calling on competitors do not have overlapping members. It will be very hard for team members to not be influenced by what has been discussed as part of developing a PSA for a competitor of the firm with which they are working. Given the current spotlight on business ethics, it is important to reach agreement on what data to share and there is a fine line between using process knowledge gained versus using competitive marketing knowledge gained from a customer or supplier. Firm employees outside of the team might execute parts of the process, but the team still maintains managerial control.

Customer relationship management and supplier relationship management are the key processes for linking firms across the supply chain and each of the other six processes is coordinated through this linkage. For example, if the customer relationship management and supplier relationship management teams decide that there is an opportunity to improve performance by focusing on the demand management process, the demand management process teams from the two companies are involved. When the process is improved, product availability is improved. If this is important, revenue for the customer increases. In addition, inventories are reduced, thereby reducing the inventory carrying cost charged to the customer's profitability report. There also may be fewer last minute production changes and less expediting of inbound materials which will impact the costs assigned to each customer. It is important that metrics are in place for the demand management process teams so that members can be compensated for the improvements derived. However, if profitability reports by customer are properly developed, they will capture improvements made in all of the processes. So having accurate profitability reports is key.

23.10 Measuring the Financial Impact of Customer Relationship Management and Supplier Relationship Management

The development of customer profitability reports enables the customer relationship management process teams to track performance over time. These reports should reflect all of the cost and revenue implications of the relationship. Variable manufacturing costs are deducted from net sales to calculate a manufacturing contribution. Next, variable marketing and logistics costs, such as sales commissions, transportation, warehouse handling, special packaging, order processing and a charge for accounts receivable, are deducted to calculate a contribution margin. Assignable nonvariable costs, such as salaries, customer related advertising expenditures, slotting allowances and inventory carrying costs, are subtracted to obtain a segment controllable margin. The net margin is obtained after deducting a charge for dedicated assets. Because these statements contain opportunity costs for investments in receivables and inventory and a charge for dedicated assets, they are much closer to cash flow statements than traditional profit and loss statements. They contain revenues minus the costs (avoidable costs) that disappear if the revenue disappears.

Sysco, a $23.4 billion food distributor, implemented profitability reports by customer in 1999. These reports enabled management to make strategic decisions about the allocation of resources to accounts, such as which customers receive preferred delivery times, which customers receive value-added services free and which ones must pay for them. The result was increased profit growth as shown in Figure 23.6. The five-year cumulative annual growth rate for the period 1999–2003 was 11.3% for sales and 19.1% for net earnings. As shown in Figure 23.6, the rate of growth in net earnings improved sharply after the profitability reports were implemented.

In the case of retailers and wholesalers, profitability reports also can be developed for each supplier. However, for manufacturers who purchase materials, total cost reports are used to evaluate suppliers. In addition to measuring current performance, these profitability reports and total cost reports can

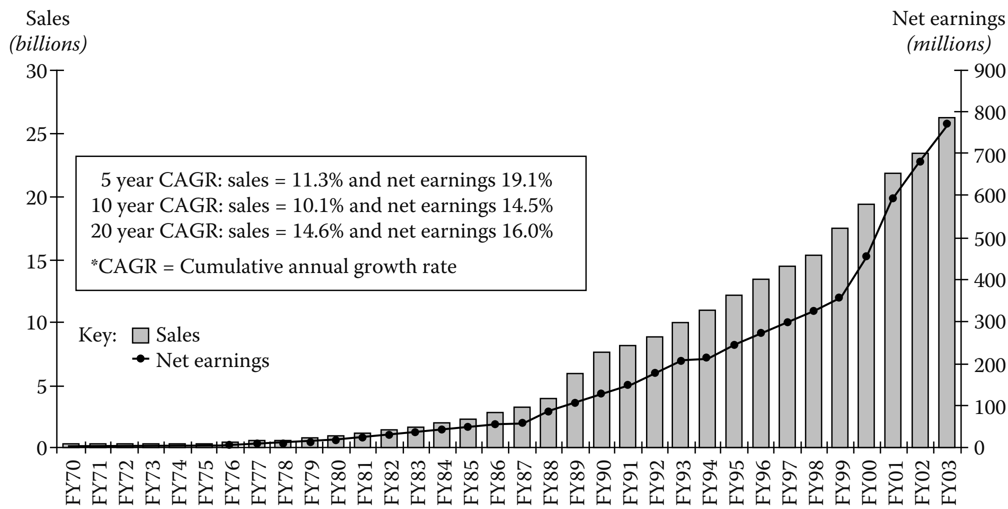


FIGURE 23.6 Sysco sales and earnings history. (From Neil Theiss, Senior Director, Supply Chain Management, Sysco Corporation. With permission.)

be used to track performance of customers and suppliers over time and to generate pro-forma statements that can be used to evaluate potential process improvement projects. Decision analysis can be performed to consider "what if" scenarios such as best, worst and most likely cases.

Figure 23.7 shows how the customer relationship management process can affect the firm's financial performance as measured by economic value added (EVA®) [53]. It illustrates how customer relationship management can impact sales, cost of goods sold, total expenses, inventory investment, other current assets, and the investment in fixed assets. For example, customer relationship management can lead to higher sales volume as a result of strengthening relationships with profitable customers, selling higher margin products, increasing the firm's share of the customer's expenditures for the products/services sold, and/or improving the mix, that is, aligning services and the costs to serve. The same approach can be used for each of the eight SCM processes to measure its impact on EVA®.

Management should implement processes that increase the profitability of the total supply chain not just the profitability of a single firm. Implementing SCM should benefit the whole supply chain while members share equitably in the risks and the rewards. If the management team of a firm makes a decision that positively affects that firm's EVA® at the expense of the EVA® of customers or suppliers, every effort should be made to share the benefits in a manner that improves the financial performance of each firm involved and thus give each one an incentive to improve overall supply chain performance.

23.11 Building High-performance Relationships in the Supply Chain

Successful implementation of GSCF SCM Framework is dependent on developing close relationships with key customers and suppliers. In other words, supply chain management is relationship management. For this reason, there is a need for a tool that can be used to structure the key relationships that are identified when implementing customer relationship management and supplier relationship management. This tool is the partnership model and the GSCF definition of partnership follows:

> A partnership is a *tailored* business relationship based on mutual trust, openness, shared risk and shared rewards that results in business performance greater than would be achieved by the two firms working together in the absence of partnership [54].

Partnerships can take multiple forms and the degree of partnership achieved can reflect tight integration across the firm boundaries, or only limited integration across the boundaries. Since partnership implementation requires significant managerial time commitments and often other resource commitments, the goal is to fit the type of partnership to the business situation and the organizational environment. The types of partnership are Type I, Type II and Type III. These are called "types", not "levels" because there should be no implication that higher levels are better than lower levels. The goal should be to have the correct amount of partnering in the relationship. Figure 23.8 shows the range of possible relationships.

23.12 The Partnership Model

The model separates the drivers of partnership, the facilitators of partnership, the components of partnership and the outcomes of partnership into four major areas for attention (see Figure 23.9). Drivers are the compelling reasons to partner, and must be examined first when approaching a potential partner. Facilitators are characteristics of the two firms that will help or hinder the partnership development process. Components are the managerially controllable elements that should be implemented at a particular level depending on the type of partnership. Outcomes measure the extent to which each firm achieves its drivers. The partnership model provides a structure for assessing the drivers and facilitators, and component descriptions for the prescribed type of partnership.

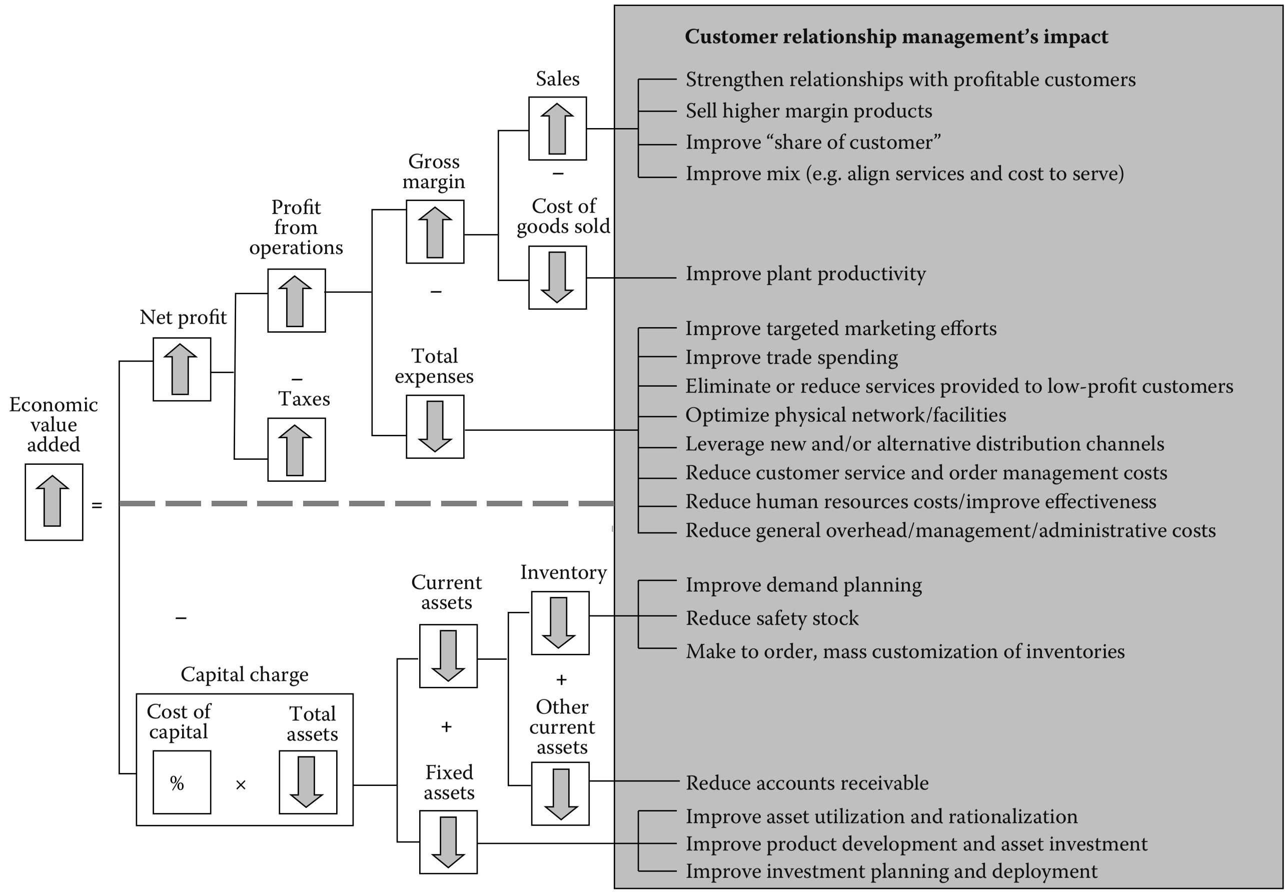


FIGURE 23.7 How customer relationship management affects economic value added (EVA®). (Adapted from Douglas M. Lambert, and Terrance L. Pohlen. 2001. Supply chain metrics. *The International Journal of Logistics Management*, 12(1), 10.)

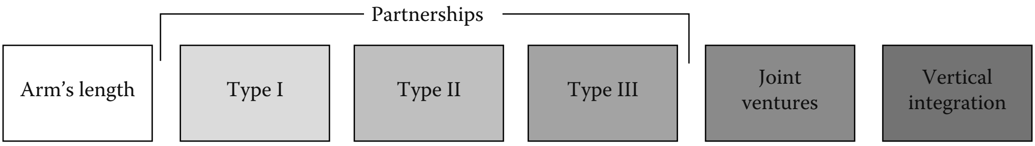


FIGURE 23.8 Types of relationships. (From Douglas M. Lambert, Margaret A. Emmelhainz, and John T. Gardner. 1996. Developing and implementing supply chain partnerships. *The International Journal of Logistics Management*, 7(2), 2. With permission.)

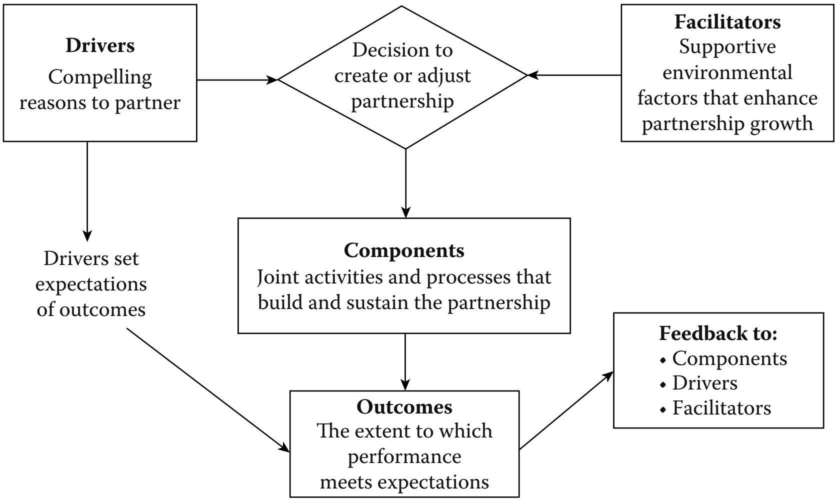


FIGURE 23.9 Partnership model. (From Douglas M. Lambert, Margaret A. Emmelhainz, and John T. Gardner. 1996. Developing and implementing supply partnerships. *The International Journal of Logistics Management*, 7(2), 4. With permission.)

23.12.1 Drivers

Why add managerial complexity and commit resources to a supply chain relationship if a good, long-term contract that is well specified will do? To the degree that business as usual will not get the supply chain efficiencies needed, partnership may be necessary. By looking for compelling reasons to partner, the drivers of partnership, management in the two firms may find that they both have an interest in tailoring the relationship. The model separates the drivers into four categories: asset/cost efficiencies, customer service improvements, marketing advantage, and profit stability and growth. All businesses are concerned with these four issues, and the four can capture the goals of managers for their relationships.

23.12.2 Facilitators

The nature of the two firms involved in partnership implementation will determine how easy or hard it will be to tailor the relationship. If the two firms mesh easily, the managerial effort and resources devoted to putting the correct relationship in place will be lower for the same results. The elements that make partnership implementation easy or hard are called facilitators. They represent the environment of the partnership; those aspects of the two firms that will help or hinder partnership activities. There are four major categories of facilitators: corporate compatibility, management philosophy and techniques, mutuality and symmetry.

23.12.3 Components

While drivers and facilitators determine the potential for partnership, the components are the building blocks of partnership. They are universal across firms and across business environments and unlike drivers and facilitators, are under the direct control of the managers involved. In other words, they are the activities that managers in the two firms actually perform to implement the partnership. There are eight components of partnership: planning, joint operating controls, communications, risk/reward sharing, trust and commitment, contract style, scope and investment. The components are implemented differently for Type l, Type II and Type III partnerships. Action items are identified for the drivers and components so that both partners' expectations are met.

23.12.4 Outcomes

A partnership, if appropriately established and effectively managed, should improve performance for both parties. Profit enhancement, process improvements, and increased competitive advantage are all likely outcomes of effective partnerships. Specific outcomes will vary depending upon the drivers which initially motivated the development of the partnership. It should be noted, however, that a partnership is not required to achieve satisfactory outcomes from a relationship. Typically, organizations will have multiple arm's length relationships which meet the needs of and provide benefits to both parties.

23.12.5 The Partnership Building Session

Using the partnership model to tailor a relationship requires a one and one-half day session. The correct team from each firm must be identified and committed to a meeting time. These teams should include top managers, middle managers, operations personnel and staff personnel. A broad mix, both in terms of management level and functional expertise, is required in order to ensure that all perspectives are considered.

The success of the partnership building process depends on the openness and creativity brought to the session. The process is not about whether to have a business relationship; it is about the style of the relationship. The partnership building session is only a first step in a challenging but rewarding long-term effort to tailor your business relationship for enhanced results.

23.13 Summary of the Supply Chain Management Framework

Figure 23.10 illustrates the interrelated nature of SCM and the need to proceed through several steps to design and successfully manage a supply chain. The SCM framework consists of three closely

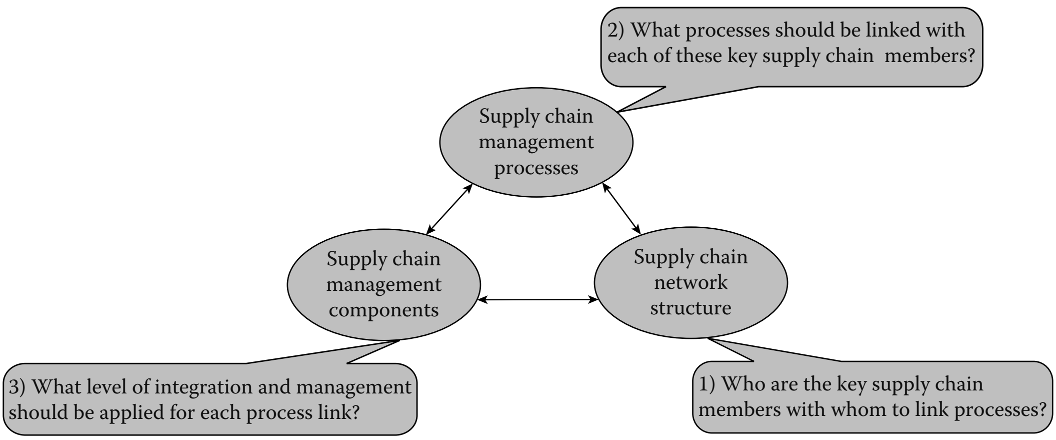


FIGURE 23.10 Supply chain management: Elements and key decisions. (Adapted from Douglas M. Lambert, Martha C. Cooper, and Janus Pagh. 1998. Supply chain management: Implementation issues and research opportunities. *The International Journal of Logistics Management*, 9(2), 4.)

interrelated elements: the supply chain network structure, the SCM processes, and the SCM components. The supply chain network structure is comprised of the member firms and the links between these firms. Business processes are the activities that produce a specific output of value to the customer. The SCM components are the managerial methods by which the business processes are integrated and managed across the supply chain. These topics are covered in detail in reference [1].

23.14 Conclusions

Executives are becoming aware of the emerging paradigm of internetwork competition, and that the successful integration and management of the SCM processes across members of the supply chain will determine the ultimate success of the single enterprise. Organizations exist in supply chains whether the relationships are managed or not. Managing the supply chain cannot be left to chance.

Research with member firms of GSCF indicates that successful SCM requires integrating business processes with key members of the supply chain. Considerable waste of valuable resources results when supply chains are not integrated, appropriately streamlined and managed. The structure of activities/processes within and between companies is vital for creating superior competitiveness and profitability. A prerequisite for successful SCM is to coordinate activities within the firm by implementing the eight SCM processes using cross-functional teams. The partnership model is a tool that can be used to structure these cross-functional relationships with key customers and suppliers.

Failure to implement cross-functional business processes will result in missed opportunities that with the level of competitiveness faced by most firms can no longer be tolerated. For example, a manufacturer of consumer durable goods implemented a rapid delivery system that provided retailers with deliveries in 24 or 48 hours anywhere in the United States. The rapid delivery system was designed to enable the retailers to improve service to retail consumers while holding less inventory and thus improving per unit profitability. Six years later, the company had not seen the anticipated reductions in retailers' inventories and reduced the service promise to 48 or 72 hours depending on the retailer's location. The rapid delivery system never achieved its full potential because the sales and marketing organizations still provided customers with incentives to buy in large volumes [55].

This example should make it clear that failure to manage all the touches will diminish the impact of initiatives within the supply chain. Implementing the eight supply chain management processes will increase the likelihood of success because all functions as well as key customers and suppliers will be involved in the planning and implementation of the initiative. The penalty for not gaining the full involvement of all functions and aligning the metrics is dealing with the actions of those who maliciously or inadvertently undermine the initiatives.

The implementation of SCM involves identifying: the supply chain members, with whom it is critical to link; the processes that need to be linked with each of these key members; and, the type/level of integration that applies to each process link. The objective of SCM is to create the most value not simply for the company but the whole supply chain network including the end-customer. Consequently, supply chain process integration and reengineering initiatives should be aimed at boosting total process efficiency and effectiveness across members of the supply chain.

At a meeting of GSCF, a series of break-out sessions were devoted to the topic "the supply chain of the future". At the end of the day, the conclusion of the group was that when an organization's management had successfully implemented all eight of the SCM processes, they would have achieved the supply chain of the future and would be able to respond to whatever challenges the business might face. Where is your company in terms of successful implementation of cross-functional business processes? In order to create the most value for the company's shareholders and the whole supply chain including end users/consumers, management must take action to integrate the supply chain.

References

1. Lambert, Douglas M., Editor. 2008. *Supply Chain Management: Processes, Partnerships, Performance*, 3rd edition. Sarasota, FL: Supply Chain Management Institute.
2. Simchi-Levy, David, Philip Kaminski, and Edith Simchi-Levy. 2000. *Designing and Managing the Supply Chain: Concepts, Strategies, and Case Studies*. Boston, MA: Irwin/McGraw Hill.
3. Monczka, Robert M., Robert J. Trent, and Robert B. Handfield. 1988. *Purchasing and Supply Chain Management*. Cincinnati, OH: South-Western College Publishing.
4. Wisner, Joel D., G. Keong Leong, and Keah-Choon Tan. 2004. *Supply Chain Management: A Balanced Approach*. Mason, OH: Thomson South-Western.
5. Drucker, Peter F. 1998. Management's new paradigms. *Forbes Magazine*, October 5, 152–177; and Christopher, Martin G. 1998. Relationships and alliances: embracing the era of network competition. In *Strategic Supply Chain Management*. John Gattorna, Editor. Hampshire, England: Gower Press, 272–284.
6. Lambert, Douglas M., Martha C. Cooper, and Janus D. Pagh. 1998. Supply chain management: implementation issues and research opportunities. *The International Journal of Logistics Management*, 9(2), 1–19.
7. The Global Supply Chain Forum, Fisher College of Business, The Ohio State University. See: fisher.osu.edu/scm.
8. Oliver, R. Keith, and Michael D. Webber. 1982. Supply-chain management: Logistics catches up with strategy. *Outlook*, cit. Christopher, Martin G. 1992. *Logistics, The Strategic Issue*. London: Chapman and Hall.
9. La Londe, Bernard J. 1998. Supply chain evolution by the numbers. *Supply Chain Management Review*, 2(1), 7–8.
10. Stevens, Graham C. 1989. Integration of the supply chain. *International Journal of Physical Distribution and Logistics Management*, 19(8) 3–8; and Towill, Denis R., Mohamed M. Naim, and J. Wikner. 1992. Industrial dynamics simulation models in the design of supply chains. *International Journal of Physical Distribution and Logistics Management*, 22(5), 3–13.
11. Bechtel, Christian, and Jayanth Jayaram. 1997. Supply chain management: A strategic perspective. *The International Journal of Logistics Management*, 8(1), 15–34.
12. Davis, Tom. 1993. Effective supply chain management. *Sloan Management Review*, 34(4), 35–46; Arntzen, Bruce C., Gerald G. Brown, Thomas P. Harrison, and Linda L. Trafton. 1995. Global supply chain management digital equipment corporation. *Interfaces*, 25(1), 69–93; and, Camp, Robert C., and Dan N. Colbert. 1997. The Xerox quest for supply chain excellence. *Supply Chain Management Review*, Spring, 82–91.
13. Scharlacken, John W. 1998. The seven pillars of global supply chain planning. *Supply Chain Management Review*, 2(1), 32–40; Tyndall, Gene, Christopher Gopal, Wolfgang Partsch, and John Kamauff. 1998. *Supercharging Supply Chains*. New York: John Wiley & Sons; and, Copacino, William C. 1997. *Supply Chain Management: The Basics and Beyond.* Boca Raton, FL: St. Lucie Press.
14. Fisher, Marshall L. 1997. What is the right supply chain for your product? *Harvard Business Review*, 75(2), 105–116; Hau L. Lee, and Corey Billington. 1992. Managing supply chain inventory: pitfalls and opportunities. *Sloan Management Review*, 33(3), 65–73; Handfield, Robert B., and Ernest L. Nichols, Jr. 1999. *Introduction to Supply Chain Management*. Upper Saddle River, NJ: Prentice Hall; and, Bowersox, Donald J. and David J. Closs. 1996. *Logistical Management–The Integrated Supply Chain Process*. New York, NY: McGraw-Hill.
15. CLM defined logistics management as: The process of planning, implementing, and controlling the efficient, cost-effective flow and storage of raw materials, in-process inventory, finished goods, and related information flow from point-of-origin to point-of-consumption for the purpose of conforming to customer requirements. *What's It All About?* Oak Brook, IL: Council of Logistics Management, 1986.

16. The definition is posted at the CSCMP's homepage: http://www.CSCMP.org (Accessed:2008).
17. APICS. 2001. *Basics of Supply Chain Management*. CPIM Certification Review Course, Participant Guide, Version 2.1, Alexandria, VA: APICS, The Educational Society for Resource Management.
18. Monczka, Robert M., Robert J. Trent, and Robert B. Handfield. 1998. *Purchasing and Supply Chain Management*. Cincinnati, OH: South-Western College Publishing.
19. Wisner, Joel D., G. Keong Leong and Keah-Choon Tan. 2004. *Supply Chain Management: Balanced Approach*. Mason, OH: Thomson South-Western.
20. Blackstock, Thomas. 2005. Keynote Speech, International Association of Food Industry Suppliers, San Francisco, CA, March 11, 2005.
21. Gattorna, John. 2006. Supply chains are the business. *Supply Chain Management Review*, 10(6), 42–49.
22. This section is based on material included in Lambert, Douglas M., Sebastián J. García-Dastugue, and Keely L. Croxton. 2005. An evaluation of process-oriented supply chain management frameworks. *Journal of Business Logistics*, 26(1), 25–51.
23. Davenport, Thomas H., and Michael C. Beers. 1995. Managing information about processes. *Journal of Management Information Systems*, 12(1), 57–80.
24. Hammer, Michael. 2001. The superefficient company. *Harvard Business Review*, 79(8), 82–91.
25. Hammer, Michael, and Glenn E. Mangurian. 1987. The changing value of communications technology. *Sloan Management Review*, 28(2), 65–71; and, Davenport, Thomas H., Michael Hammer and Tauno J. Metsisto. 1989. How executives can shape their company's information systems. *Harvard Business Review*, 67(2), 130–134.
26. Hammer, Michael, and James Champy. 1993. *Reengineering the Corporation: A Manifesto for Business Revolution*, 1st edition. New York, NY: Harper Business.
27. Davenport, Thomas H. 1993. *Process Innovation: Reengineering Work through Information Technology*. Boston, MA: Harvard Business School Press.
28. Hammer, Michael and Glenn E. Mangurian. 1987. The changing value of communications technology. *Sloan Management Review*, 28(2), 65–71.
29. Hammer, Michael. 1990. Reengineering work: Don't automate, obliterate. *Harvard Business Review*, 68(4), 104–112.
30. Taylor, Frederick W. 1911. *The Principles of Scientific Management*. New York, NY: Harper & Bros.
31. Davenport, Thomas H., and James E. Short. 1990. The new industrial engineering: information technology and business process redesign. *Sloan Management Review*, 31(4), 11–27.
32. Weld, Louis D. H. 1917. Marketing functions and mercantile organizations. *The American Economic Review*, 7(2), 306–318.
33. Kotler, Philip. 1991. *Marketing Management: Analysis, Planning, Implementation and Control*, 7th edition. Englewood Cliffs, NJ: Prentice-Hall.
34. McCarthy, E. Jerome. 1960. *Basic Marketing: A Managerial Approach*. Homewood, IL: R.D. Irwin.
35. Alderson, Wroe. 1950. Marketing efficiency and the principle of postponement. *Cost and Profit Outlook*, 3, September; Reavis Cox and Wroe Alderson, Editors. 1950. *Theory in Marketing*. Chicago, IL: Richard D. Irwin, Inc.; and, Bucklin, Louis P. 1966. *A Theory of Distribution Channel Structure*. IBER Special Publication, Berkeley, CA.
36. Webster, Frederick E. Jr. 1992. The changing role of marketing in the corporation. *Journal of Marketing*, 56(4), 1–17; Atul Parvatiyar and Jagdish N. Sheth. 1994. *Relationship Marketing: Theory, Methods, and Applications*, Jagdish N. Sheth and Atul Parvatiyar, Editors. Atlanta, GA: Emory University Center for Relationship Marketing.
37. Grönroos, Christian. 1994. Quo Vadis, marketing? Toward a relationship marketing paradigm. *Journal of Marketing Management*, 10(5), 347–360.
38. Grönroos, Christian. 1990. *Service Management and Marketing, Managing the Moments of Truth in the Service Competition*. Lexington, MA: Free Press/Lexington Books.
39. Webster, Frederick E. Jr. 1992. The changing role of marketing in the corporation. *Journal of Marketing*, 56(4), 1–17.

40. Webster, Frederick E. Jr. 1992. The changing role of marketing in the corporation. *Journal of Marketing*, 56(4), 1–17.
41. Cooper, Martha C., Douglas M. Lambert and Janus D. Pagh. 1997. Supply chain management: more than a new name for logistics. *The International Journal of Logistics Management*, 8(1), 1–14; Lambert, Douglas M., Martha C. Cooper and Janus D. Pagh. 1998. Supply chain management: implementation issues and research opportunities. *The International Journal of Logistics Management*, 9(2), 1–19; and, Croxton, Keely L., Sebastián J. García-Dastugue, Douglas M. Lambert and Dale S. Rogers. 2001. The supply chain management processes. *The International Journal of Logistics Management*, 12(2), 13–36.
42. Håkansson, Håkan and Ivan Snehota. 1995. *Developing Relationships in Business Networks*. London: Routledge.
43. Blackstock, Thomas. 2005. Keynote Speech, International Association of Food Industry Suppliers, San Francisco, CA, March 11, 2005.
44. Blackstock, Thomas. 2005. Keynote Speech, International Association of Food Industry Suppliers, San Francisco, CA, March 11, 2005.
45. Lambert, Douglas M. and Terrance L. Pohlen. 2001. Supply chain metrics. *The International Journal of Logistics Management*, 12(1), 1–19.
46. Bowersox, Donald J. 1997. Integrated supply chain management; a strategic perspective. *Annual Conference Proceedings*, Chicago, IL: Council of Logistics Management, 181–189; Stigler, George E. 1951. The division of labor is limited by the extent of the market. *Journal of Political Economy*, 59(3), 185–193; and, Coase, R. H. 1937. The nature of the firm. *Economica*, 4, 386–405.
47. Håkansson, Håkan and Ivan Snehota. 1995. *Developing Relationships in Business Networks*. London: Routledge.
48. Lambert, Douglas M. and A. Michael Knemeyer. 2004. We're in this together. *Harvard Business Review*, 82(12), 96–108; and Lambert, Douglas M., Margaret A. Emmelhainz, and John T. Gardner. 1996. Developing and implementing supply chain partnerships. *The International Journal of Logistics Management*, 7(2), 1–17.
49. Cooper, Martha C. and John T. Gardner. 1993. Good business relationships: More than just partnerships or strategic alliances. *International Journal of Physical Distribution and Logistics Management*, 23(6), 14–20.
50. Berry, Leonard L. and A. Parasuraman. 1991. Marketing to existing customers, In *Marketing Services: Competing Through Quality*. New York, NY: The Free Press, 132.
51. Heskett, James L., W. Earl Sasser, Fr. and Leonard A. Schlesinger. 1997. *The Service Profit Chain*. New York, NY: The Free Press, 11.
52. Seybold, Patrica B. 2001. Get inside the lives of your customers. *Harvard Business Review*, 78(5), 81–89.
53. Bennett Stewart III, G. 1999. *The Quest for Value*. New York: Harper Collins Publishers.
54. Lambert, Douglas M., and A. Michael Knemeyer. 2004. We're in this together. *Harvard Business Review*, 82(12), 114–122; and Lambert, Douglas M. 206. *Supply Chain Management; Processes, Partnerships, Performance*. Sarasota, FL: Supply Chain Management Institute, 169.
55. Lambert, Douglas M. and Renan Burduroglu. 2000. Measuring and selling the value of logistics. *The International Journal of Logistics Management*, 11(1), 1–17.

24

Hierarchical Dynamic Decision Making

Adedeji B. Badiru
Air Force Institute of Technology

24.1 Introduction

The success of manufacturing depends on timely and effective information. Planning in a manufacturing environment is influenced by the integrity of the available information. To generate such information often requires the use of forecasting. Forecasting, itself is a function of the efficacy of input data, which decision makers usually don't have. Consequently, decision makers resort to guesses based on experience and intuition. The use of such unempirical data can be enhanced by a simulation approach. The problem addressed in this chapter involves the development of a hybrid analytic dynamic forecasting methodology that combines the techniques of analytic hierarchy process (AHP), factor analysis (FA), spanning tree (ST) technique, computer simulation, and the decision maker's experiential intuition. The methodology presents a simulated scenario output to the decision maker, from which a more integrative forecast decision can be made. Military decision-making environment provides a suitable application platform for the DDM methodology. Forecast results can be used to make operations planning decisions in manufacturing. This chapter presents a probability extension and simulation implementation of AHP incorporating a ST approach and FA based on the previous work of Badiru et al. (1993) on the development of a simulation-based forecasting methodology for hierarchical dynamic decision making (DDM). The problem is to predict the level of change in a base forecast (generated by conventional techniques) due to some probabilistic qualitative factors that are not represented in the calculation of the base forecast. The ST creates a scenario set which requires fewer pair-wise comparison matrices from the decision maker. The approach generates a distribution of forecast outcomes rather than a single forecast value. Forecast scenarios are generated by rudimentary probability information specified by the decision maker. The DDM approach facilitates more interactive, data-efficient, and faster decision making under uncertainty.

24.2 Decision Making Literature Review

The AHP (Saaty 1980) facilitates the incorporation of qualitative considerations into quantitative factors for decision making. AHP is a practical approach to solving complex decision problems involving the comparison of alternatives. The technique has been used extensively in practice to address decisions under choices. Zahedi (1986), Golden et al. (1989), and Vargas (1990) present comprehensive surveys of the AHP technique and its various applications. Wind and Saaty (1980) present an example of the application of AHP to marketing. Saaty and Vargas (1979) discuss the estimation of technological coefficients by the AHP. Ramanujam and Saaty (1981) present a detailed AHP approach to the assessment and selection of imported technology in less developed countries. Madu (1988) points out the potential effectiveness of AHP in ranking different technologies for possible adoption by specific developing countries. Wabalickis (1988) presents the use of AHP for the justification of flexible manufacturing system. Banai-Kashani (1985, 1989) presents an application of AHP to urban housing and transportation planning. Rahman and Frair (1984) present an application of AHP to electric utility planning. The problem of long-range planning using AHP was addressed by Emshoff and Saaty (1982). Mustafa (1989) applies AHP to the problem of strategic planning in industry. Azani and Khorramshahgol (1990) and Khorramshahgol and Moustakis (1988) present the Analytic Delphi Method (ADM), which integrates conventional Delphi method of forecasting and AHP for location planning. Khorramshahgol et al. (1988) use AHP for project evaluation and selection. Cook et al. (1984) present an extension of the AHP technique for urban economic forecasting.

The decision hierarchy in AHP is constructed so that elements at the same level are of the same class and must be capable of being related to some elements in the next higher level. The top level of the hierarchy reflects the overall objective or focus of the decision problem. Criteria, factors, or attributes on which the final objective is dependent are listed at intermediate levels in the hierarchy. The lowest level in the hierarchy contains the competing alternatives through which the final objective might be achieved. After the hierarchy has been constructed, the decision maker must undertake subjective prioritization to determine the weight of each element at each level of the hierarchy. Pair-wise comparisons are performed at each level to determine the relative importance of each element at that level with respect to each element at the next higher level in the hierarchy. Mathematical basis, procedures, and extensions of AHP are widely available in the literature. Several issues on AHP theory and applications are addressed by Saaty (1977, 1980, 1986, 1990), Saaty and Vargas (1987), Saaty and Vargas (1991), Liberatore (1987), Saaty and Alexander (1989), Winkler (1990), Dyer (1990), Harker and Vargas (1987), Vargas (1983), Saaty and Vargas (1984), Saaty et al. (1983), Girgis et al. (1989), Masuda (1990), Forman et al. (1983), and Golden et al. (1989).

The methodology presented can also be extended to incorporate FA into AHP to develop cross-correlation effects of variables affecting a forecast outcome. FA is a statistical procedure for estimating the minimum number of separate variables or factors necessary to provide the information contained in a correlation matrix. It has been widely used in psychology and the social sciences for decades (Spearman 1904; Cattell 1978; Park et al. 2002; Russell 2002; Snook and Gorsuch 1989, Kline 2000). But its use in general engineering or management exploratory research remains largely unexplored. It is a technique that is often a necessity in some branches of psychology, especially those in which tests or qualitative surveys are used as the research basis (Kline 2000). Shapiro et al. (2002) present a good example of the application of FA on survey data from Gulf War veterans to predict possible health effects of war deployment. Koong and Liu (1999) used FA to study the attributes affecting the use of computer information systems for decision support. Blind et al. (2001) applied FA to a technology forecasting problem.

24.3 Forecast Scenario Planning

Conventional forecasting is based on static and certain relationships among the factors in the forecast problem. This results in single-value predictions. Scenario planning, by contrast, focuses on uncertainties in the forecast environment. Scenario planning generates probabilistic representations of the future.

Figure 24.1 illustrates the conventional forecasting process, in which past events and current events are used to generate static representations of future events. Figure 24.2 illustrates the scenario planning, in which past events, current events, and simulated future events are used to generate probabilistic representations of future events.

24.4 Modeling Notations

M_i, Main Event *i*.
E_{iu}, Sub-event *u* of Event *i*.
P_{iu}, Probability that E_{iu} will occur.
w_{ab}^{iu}, Likelihood of Outcome *a* over Outcome *b* with respect to E_{iu}.
d_{ij}, Overall significance of M_i over M_j.
c_{uv}^{ij}, Relative significance of E_{iu} over E_{jv}.
E'_{iu}, Sub-event *u* of the event in the *i*th position of the *rearranged sequence* of events.
c'^{ij}_{uv}, Relative Significance of E'_{iu} over E'_{jv}.
x_a^{iu}, Relative weight of outcome *a* with respect to E_{iu}.
S_k, *k*th scenario.
$e_{iu}(S_k)$ Relative weight of E_{iu} calculated with respect to S_k.
z_{iu}, Ratio of relative weight of E_{iu} to relative weight of root node, $E_{12}(z_{12}=1)$.
$C(a)$, Composite weight for *outcome a*.
Definitions: Main Event = Factor; Sub-event = Factor level; Alternative = Forecast Outcome.

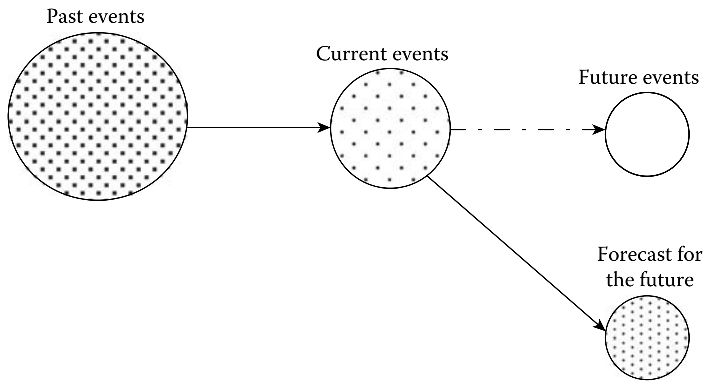


FIGURE 24.1 Conventional forecasting.

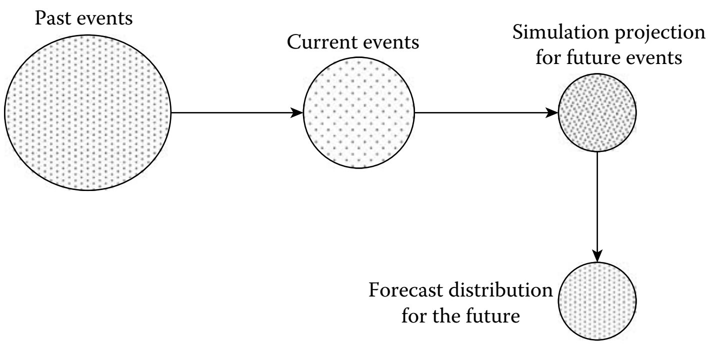


FIGURE 24.2 Forecast scenario planning.

24.5 Modeling Approach

The objective is to predict the change in the base forecast as a function of qualitative factors, which were not included in the forecast. The model consists of sets of independent events, dependent events, levels at which the events may occur, and outcomes defined as the possible variations in base forecast due to the occurrence of such events. The occurrence probabilities for the independent events and the conditional probabilities for the dependent events are provided by the decision maker. The decision maker also provides pair-wise comparisons for the set of events generated by the procedure. Sub-events in a scenario occur according to event probability distributions and the scenario is evaluated using the AHP approach. The AHP approach consists of defining a hierarchy and analyzing the elements of each level in terms of their impact on the elements at the higher level by means of pair-wise comparisons. The pair-wise comparisons are combined into an overall weight for each element. The weights are then aggregated to derive a set of ratings for the forecast outcomes. The outcome with the highest rating is selected.

Figure 24.3 presents a general model of the AHP problem. The highest hierarchy specifies the objective of the problem, which is to estimate changes in a base forecast. The second level contains a set of main events that affect the objective. The main events may be probabilistic and/or conditional in nature. For example, Event 3 depends on Event 2 and Event 4 depends on Event 3. Each main event has a certain number of sub-events. The number of sub-events does not have to be equal for all main events. Factors are main events, factor levels are sub-events, and alternatives are referred to as forecast outcome.

We need to assign probabilities to each sub-event. Let P_{iu} denote the probability that Sub-event u of Event i will occur. That is,

$$P(E_{iu}) = P_{iu}, \ \forall \ E_{iu} \tag{24.1}$$

where E_{iu} denotes Sub-event u of Event i. We will let M_i denote Main Event i. The occurrence of an event may be dependent on another event. In our example, events 1 and 2 are independent and events 3 and 4 are dependent events. We incorporate conditional probabilities, $P(E_{jv}|E_{iu})$, for all dependent events i, j and their respective sub-events u, v. The next step is pair-wise comparison of outcomes (level 4) for each E_{iu}.

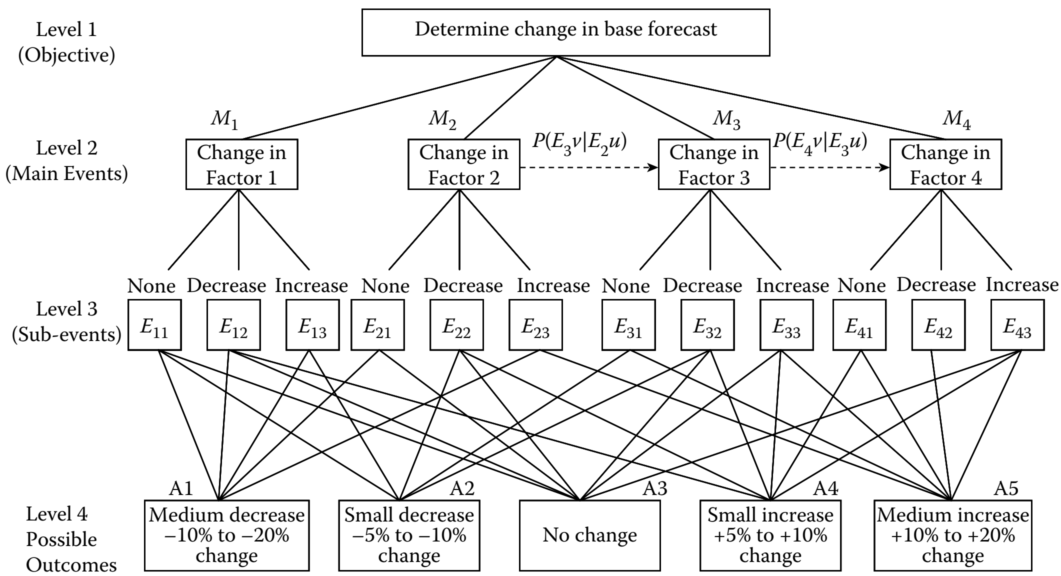


FIGURE 24.3 General model of probabilistic AHP for forecasting.

We let w_{ab}^{iu} denote the importance of Outcome *a* over Outcome *b* with respect to E_{iu}. The conventional weight scale presented by Saaty (1980) is adopted for w_{ab}^{iu}.

The hierarchical factor weights are interpreted as:

$w_{ab}^{iu} = 1$ if outcome a is equally likely as outcome *b*.
$w_{ab}^{iu} = 3$ if outcome *a* is weakly more likely than outcome *b*.
$w_{ab}^{iu} = 5$ if outcome *a* is essentially more likely than outcome *b*.
$w_{ab}^{iu} = 7$ if outcome *a* is demonstratedly more likely than outcome *b*.
$w_{ab}^{iu} = 9$ if outcome *a* is absolutely likely than outcome *b*.

Intermediate weights of 2, 4, 6, and 8 are used as appropriate. The relative importance of sub-events of Event *i* over sub-events of Event *j* are also needed for the computation of final composite weights for the outcomes. Define c_{uv}^{ij} as the relative importance of E_{iu} over E_{jv} in accordance with the weight scale described previously. It is practically impossible for the decision maker to input c_{uv}^{ij} for all pairs of *i,j* and *u,v* such that a consistent comparison matrix will be generated for each scenario. It is also practically infeasible to input the comparison matrix for each scenario separately. Our procedure requires the decision maker to provide c_{uv}^{ij} for only a subset of *i,j* and *u,v*. The remaining comparisons are internally carried out during the simulation. The procedure is as follows: First, rearrange all the main events in increasing order of the number of sub-events. This rearrangement is needed to generate the minimal scenario set. Let E'_{iu} denote sub-event *u* of the event in the *i*th position of the *rearranged sequence* of events. Let c'^{ij}_{uv} denote the comparison weight of E'_{iu} over E'_{jv}. The first sub-event of each event is defined as the event not occurring (i.e. "no change" in the parent event). Hence, c_{uv}^{ij} and c'^{ij}_{uv} are defined only for $u \neq 1$ and $v \neq 1$. Let each E'_{iu} be represented by a node. The objective is to define the minimal number of scenarios such that pair-wise comparisons of events within each scenario, if consistent, will collectively lead to consistent weight matrices for all scenarios. Although there is more than one way of generating such scenarios, our procedure adopts the heuristic presented in the following section.

24.6 Generation of Minimal Scenario Set

The procedure assumes a definite layout of nodes whereby the nodes (i.e. sub-events) belonging to the same event are arranged row-wise and the nodes corresponding to the same sub-event number are arranged column-wise. This is illustrated in Figure 24.4. Note that the main events do not necessarily appear in numerical order in the first column since they are arranged in increasing order of the number of associated sub-events.

Step 1: The first sub-event of each event is excluded from pair-wise comparisons. Thus, the first column of nodes in Figure 24.4 will not be included in the generation of scenarios. In this example, we will start marking the scenarios from node column 2. Define a scenario, S_k, as the set of sub-events such that main events occur at the same level of sub-event number. That is, all the sub-events making up S_k all have the same second subscripts. Mark all columns of S_k having two or more elements. In Figure 24.4, node columns 2, 3, and 4 satisfy Step 1 requirements. The elements in the scenarios are joined by solid arrows in the figure. Thus, there are three scenarios:

$$S_1 = \{E'_{12}, E'_{22}, E'_{32}, E'_{42}, E'_{52}\}$$

$$S_2 = \{E'_{33}, E'_{43}, E'_{53}\}$$

$$S_3 = \{E'_{44}, E'_{54}\}$$

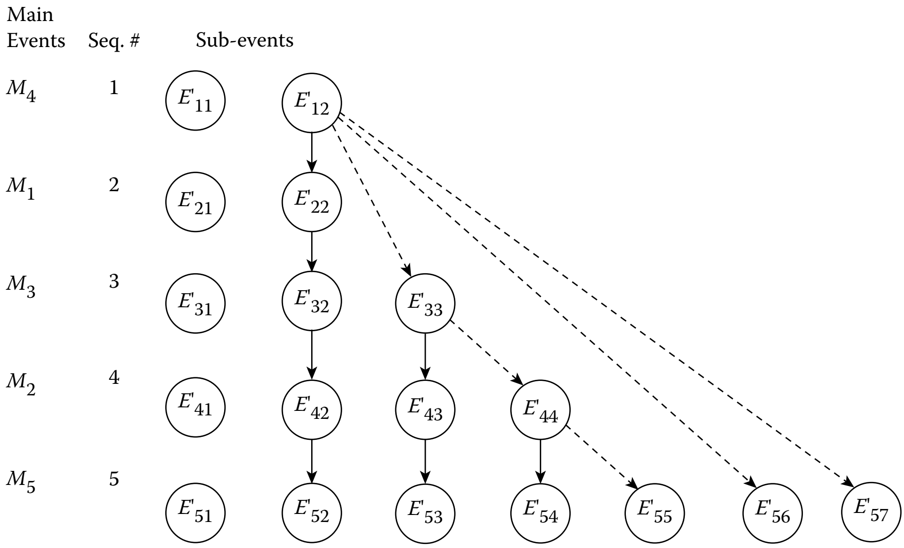


FIGURE 24.4 Generation of relative weights.

Step 2: Generate a new scenario. Let the first node in the next unmarked column be the first element of the new scenario. Include the first node in each column which has a lower sub-event number subject to the following restrictions:

(a) No sub-event, except E'_{12}, can appear in more than one scenario generated in Step 2.
(b) Not more than one sub-event of the same main event can appear in a scenario.

Add E'_{12} to the scenario if it is not already included. Stop if all columns have been marked. Otherwise, repeat Step 2. The elements of the resulting scenarios are connected by broken arrows (see Figure 24.4). The scenarios are:

$$S_4 = \left\{E'_{12}, E'_{33}, E'_{44}, E'_{55}\right\}$$

$$S_5 = \left\{E'_{12}, E'_{56}\right\}$$

$$S_6 = \left\{E'_{12}, E'_{57}\right\}$$

The arcs of the network define a ST rooted at E'_{12}. By generating a small number of scenarios (defined by the paths of the ST), one can determine the relative weights of E'_{iu} and normalize the weights for each scenario. In summary, once the hierarchy structure is determined, the decision maker needs to input the following:

1. P_{iu} for each Sub-event u of *independent* Event i.
2. $P(E_{jv}|E_{iu})$ for each Sub-event v of *dependent* Event j.
3. Comparison matrix $\mathbf{W}_{iu} = [\, w_{ab}^{iu} \,]$ of outcomes with respect to E_{iu}.
4. Comparison matrix $\mathbf{C}_{Sk} = [\, c_{uv}^{ij} \,]$ for each pair of E_{iu} and E_{jv} in S_k.

The consistency of each comparison matrix is checked by comparing its maximum eigenvalue to the number of elements in the matrix (Saaty 1980). More specifically, a comparison matrix is consistent if the ratio of $(\lambda_{max} - n)/(n-1)$ to the average random index for the same order matrix is less than 10%, where λ_{max} denotes the maximum eigenvalue of the comparison matrix and n denotes the size of the matrix. The matrices are modified until the consistency condition is satisfied. The normalized eigenvector of a consistent matrix defines the relative weights of the elements in that matrix. Let x_a^{iu} denote the relative weight of outcome a with respect to E_{iu}. Let $e_{iu}(S_k)$ be the relative weight of E_{iu} calculated with respect to scenario S_k. The procedure determines $e_{iu}(S_k)$ and x_a^{iu} using $\mathbf{C}_{S_k}$ and W_{iu} matrices, respectively. Suppose there exist K independent scenarios (i.e. S_k, $k=1, 2,\ldots, K$ generated by using the scenario generation procedure). Note that $\cup_{k=1}^{K} S_k$ contains all E_{iu} except E_{i1}. Also observe that for the illustrative example, we have:

$$\bigcap_{k=1}^{K} S_k = \left\{E'_{12}, E'_{33}, E'_{44}\right\},$$

which represents the set of lead sub-events in columns having more than one sub-event. The relative weight of E_{iu} with respect to a new $S_j \neq S_k$, $k=1, 2,\ldots, K$, can be calculated as shown below. Let

$$z_{iu} = \frac{e_{iu}(S_k)}{e_{12}(S_k)}, \forall E_{iu} \in S_k, \quad k=1,2,\ldots,K \tag{24.2}$$

where z_{iu} is the ratio of the relative weight of E_{iu} to the relative weight of the root node, E_{12}. Note that $z_{12}=1$. For a new S_j, we have

$$e_{iu}(S_j) = \frac{z_{iu}}{\sum_{E_{iu} \in S_j} z_{iu}} \tag{24.3}$$

Once the comparison matrices are filled, new scenarios are generated by using E_{iu}'s, not E'_{iu}'s. This is done internally by the simulation program. First, independent events are randomly generated using the P_{iu} values. Next, conditional probabilities are used to generate dependent events. A scenario is defined by the set of events and sub-events which are assumed to have occurred. Then, relative weight, $e_{iu}(S_j)$ of each E_{iu} and the relative weight, x_a^{iu}, of each outcome a are calculated as discussed above. It should be noted that

$$\sum_{a=1}^{N} x_a^{iu} = 1 \tag{24.4}$$

with respect to each E_{iu}, where $a=1, 2,\ldots, N$ outcomes. Finally, the composite weight, $C(a)$, for each outcome a with respect to each scenario j is calculated as:

$$C(a) = \sum_i x_a^{iu} e_{iu}(S_j), \forall E_{iu} \in S_j \tag{24.5}$$

The outcome with the highest $C(a)$ value in the current scenario is selected in this specific run of the simulation. The procedure then generates a new set of random numbers and, hence, a new scenario. For each scenario, the most likely outcome is determined. After a sufficient number of repetitions, one can determine the frequency of selection for each outcome. This frequency distribution is presented to the decision maker rather than an estimated single forecast. The example below demonstrates the implementation of the approach.

24.7 Computational Example

Figure 24.5 shows a simple structure with two possible outcomes, where events 1 and 3 are independent. This simple illustrative example is used for manual computational convenience. The occurrence of Event 2

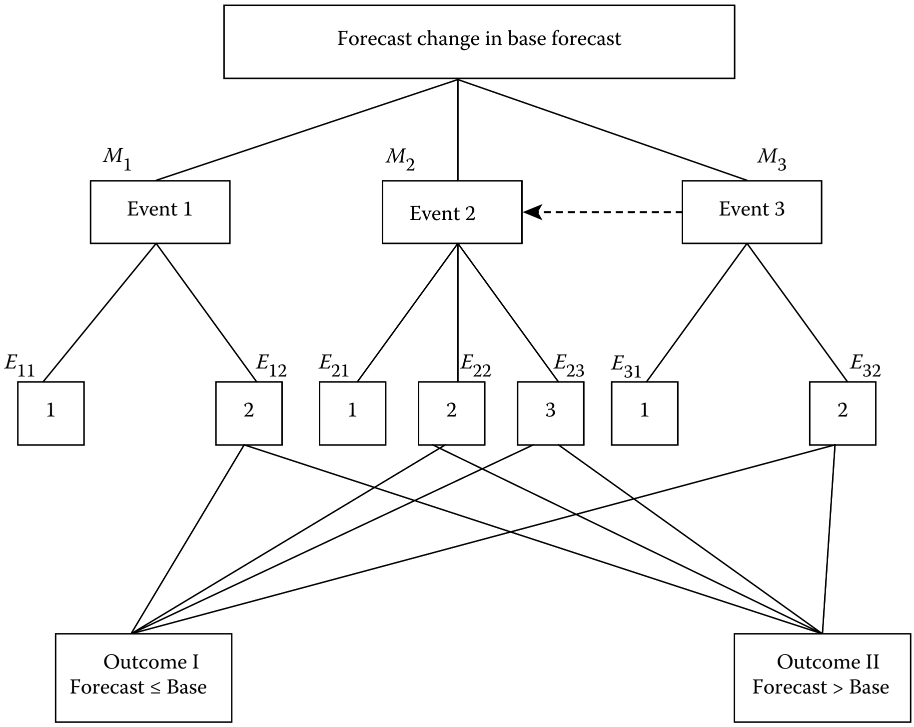


FIGURE 24.5 Example AHP network.

depends on the level at which Event 3 occurs. The first sub-event under each event describes nonoccurrence of that event. If all three events do not occur, then no outcome is picked. Main Event 1, M_1, contains two sub-events, E_{11} and E_{12}. Main Event 2, M_2, contains three sub-events, E_{21}, E_{22}, and E_{23}. Main Event 3, M_3, contains two sub-events, E_{31} and E_{32}. Reordering the events in increasing order of number of sub-events, we get the following:

$$E'_{1u} = E_{1u}, \quad u = 1,2$$

$$E'_{2u} = E_{13u}, \quad u = 1,2$$

$$E'_{3u} = E_{2u}, \quad u = 1,2,3.$$

Figure 24.6 shows the ST structure. Table 24.1 displays P_{ij} values for the independent events.

Tables 24.2 and 24.3 display $P(E_{2v}|E_{3v})$ and w_{ab}^{iu} values, respectively. Negative values are used to indicate reciprocal preference in the pair-wise comparison matrices. Table 24.3 contains four consistent pair-wise comparison matrices. The entries in Table 24.3 are used later to compute the x_a^{iu} values for the outcomes.

Table 24.4 shows the corresponding comparison matrices for the two scenarios below:

$$S_1 = \left\{E'_{12}, E'_{22}, E'_{32}\right\} \qquad S_2 = \left\{E'_{12}, E'_{33}\right\}$$

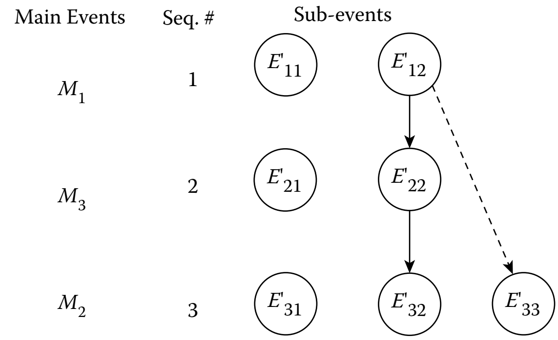


FIGURE 24.6 Spanning tree example.

TABLE 24.1 P_{iu} for Independent Events

	u	Sub-event 1	Sub-event 2
Main Event 1	P_{1u}	0.8	0.2
Main Event 2	P_{3u}	0.6	0.4

TABLE 24.2 Conditional Probabilities for Dependent Events

		E_{2v}		
	Given E_{3u}	E_{21}	E_{22}	E_{23}
Sub-event 1 of Main Event 3	E_{31}	0.6	0.3	0.1
Sub-event 2 of Main Event 3	E_{32}	0.2	0.4	0.4

TABLE 24.3 Examples of Pair-wise Comparisons of Outcomes

w^{12}_{ab}		b	w^{22}_{ab}		b	w^{23}_{ab}		b	w^{32}_{ab}		b
a	1	2	a	1	2	a	1	2	a	1	2
1	1	3	1	1	−2	1	1	5	1	1	−7
2	−3	1	2	2	1	2	−5	1	2	7	1

That is,

$$S_1 = \left\{E'_{12}, E'_{32}, E'_{22}\right\} \qquad S_2 = \left\{E'_{12}, E'_{33}\right\}.$$

The relative weights of the sub-events in scenarios 1 and 2 are calculated from the comparison matrices in Table 24.4.

$$e_{12}(S_1)=0.53,\ e_{32}(S_1)=0.33,\ e_{22}(S_1)=0.14,\ e_{12}(S_2)=0.33,\ e_{23}(S_2)=0.67$$

Using Equation 24.2, the ratios, z_{iu}, are computed for the elements in each scenario:

$$z_{12}=1,\ z_{32}=0.62,\ z_{22}=0.26,\ z_{23}=2.03$$

To generate a new scenario, a random number, r_1, is generated from a uniform distribution (0,1) for Event 1. If $r_1 \leq 0.8$, then, based on P_{11} in Table 24.1, we conclude that E_{11} occurs. Otherwise, E_{12} occurs. Similarly, random number r_2 determines level at which Event 3 occurs (E_{3u}). Suppose E_{12} and E_{32}

TABLE 24.4 Comparison Matrices for S_1 and S_2

C_{S_1}	E_{12}	E_{32}	E_{22}	C_{S_2}	E_{12}	E_{23}
E_{12}	1	3	2	E_{12}	1	−2
E_{32}	−3	1	−2	E_{23}	2	1
E_{22}	−2	2	1			

have occurred. Random number r_3 is generated and occurrence of E_{2u} (u=1, 2, 3) is determined using $P(E_{2u}|E_{32})$. Suppose that R_3=0.8. Using the second row of Table 24.2, we conclude that E_{23} has occurred. The new scenario is now defined by:

$$S_3=\{E_{12}, E_{23}, E_{32}\}$$

Using Equation 24.3, e_{iu} for each $E_{iu}\in S_3$ is calculated as:

$$e_{12}(S_3)=\frac{z_{12}}{\sum_{E_{iu}\in S_3} z_{iu}}=\frac{1}{(1+2.03+0.62)}=0.274$$

$$e_{23}(S_3)=\frac{z_{23}}{\sum_{E_{iu}\in S_3} z_{iu}}=\frac{2.03}{(1+2.03+0.62)}=0.556$$

$$e_{32}(S_3)=\frac{z_{32}}{\sum_{E_{iu}\in S_3} z_{iu}}=\frac{0.62}{(1+2.03+0.62)}=0.170$$

The relative weights x_a^{iu} for Outcomes I and II are obtained from the relevant comparison matrices based on E_{12}, E_{23}, and E_{32} in Table 24.3.

$$x_1^{12}=0.750 \quad x_1^{23}=0.833 \quad x_1^{32}=0.125$$

$$x_2^{12}=0.250 \quad x_2^{23}=0.167 \quad x_2^{32}=0.875$$

The composite weights for the two outcomes are now calculated from Equation 24.5:

$$C(1)=x_1^{12}e_{12}(S_3)+x_1^{23}e_{23}(S_3)+x_1^{32}e_{32}(S_3)$$

$$=0.750(0.274) + 0.833(0.556) + 0.125(0.170) = 0.690$$

$$C(2)=x_2^{12}e_{12}(S_3)+x_2^{23}e_{23}(S_3)+x_2^{32}e_{32}(S_3)$$

$$= 0.250(0.274) + 0.167(0.556) + 0.875(0.170) = 0.310$$

Hence, Outcome I is forecasted for this scenario. The above scenario generation and selection of outcomes are repeated n times and the frequency distribution for each outcome is determined.

24.8 Potential for Further Development

As further development, the research methodology can be extended to incorporate FA as well as analytical enhancements to the basic approach. Specific ideas for further research and development are:

- Evaluate complex probabilistic dependency structures among the event set.
- Investigate consistency issues under the scenario planning logic.
- Formulate hybrid measure of data uncertainty. Since the relevance of the results depends on the validity of the input data, a measure of uncertainty affecting the probabilities entered by the decision maker will need to be formulated into the hybrid methodology.
- Evaluate the efficiency of the approach in generating minimum number of scenarios. How do we guarantee that the scenario set is minimum rather than a conveniently small set?
- Incorporate an analytical model of FA.
- Evaluate the risks associated with forecast errors from conventional AHP forecast profiles and compare to forecast errors from the methodology.
- Evaluate time and cost to implement the methodology relative to the improved decision accuracy.

24.9 Software Implementation

The methodology was previously implemented as software implementation using C programming language (Badiru et al. 1993). Table 24.5 shows a comparison of the results of the manual computation presented in this chapter and the pilot simulation. The results, based on 15,000 simulated scenarios, indicate a close match between the pilot software implementation and analytical computations. The histogram in Figure 24.7 shows the relatives weights of the simulated forecast alternatives.

TABLE 24.5 Comparison of the Results of Manual Computation and Pilot Simulation Exercise

	Manual	Simulation
Alternative 1	0.69	0.66
Alternative 2	0.31	0.34

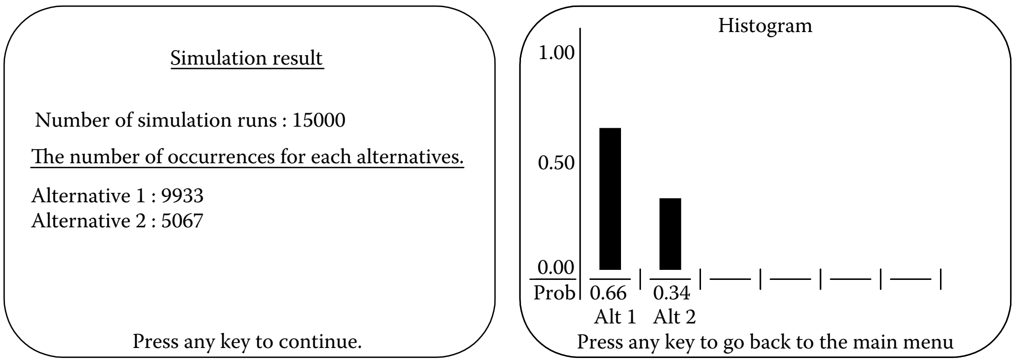


FIGURE 24.7 Histogram of simulated decision forecasts.

References

Azani, H., R. Khorramshahgol. 1990. Analytic Delphi method (ADM): A strategic decision making model applied to location planning. *Engineering Costs and Production Economics*, 20, 23–28.

Badiru, A.B., P.S. Pulat, and M. Kang. 1993. DDM: Decision support system for hierarchical dynamic decision making. *Decision Support Systems*, 10(1), 1–18.

Banai-Kashani, A.R. 1985. Building systems innovations in urban housing production: A planning application of the analytic hierarchy process. *Architectural Science Review*, 29, 99–109.

Banai-Kashani, R. 1989. Discrete mode-choice analysis of urban travel demand by the analytic hierarchy process. *Transportation*, 16, 81–96.

Blind, K., K. Cuhls, and H. Grupp. 2001. Personal attitudes in the assessment of the future of science and technology: A factor analysis approach. *Technological Forecasting and Social Change*, 68, 131–149.

Cattell, R.B. 1978. *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*. Plenum, New York.

Cook, T., P. Falchi, and R. Mariano 1984. An urban allocation model combining time series and analytic hierarchical methods. *Management Science*, 30(2), 198–208.

Dyer, J.S. 1990. Remarks on the analytic hierarchy process. *Management Science*, 36(3), 249–258.

Emshoff, J.R., and T.L. Saaty. 1982. Applications of the analytic hierarchy process to long range planning. *European Journal of Operational Research* 10, 131–143.

Forman, E.H., T.L. Saaty, M.A. Selly, and R. Waldom. 1983. *Expert Choice, Decision Support Software*. McLean, VA.

Girgis, A.A., M.A. Wortman, and R. Scott Henry. 1989. Development of rank ordering techniques based on the analytical hierarchy process for evaluation of energy conversion devices. *International Journal of Energy Research*, 13, 279–287.

Golden, B.L., E.A. Wasil, and P.T. Harker, editors. 1989. *The Analytic Hierarchy Process: Applications and Studies*. Springer-Verlag, New York.

Harker, P.T., and L.G. Vargas. 1987. The theory of ratio scale estimation: Saaty's analytic hierarchy process. *Management Science*, 33(11), 1383–1403.

Khorramshahgol, R., H. Azani, and Y. Gousty. 1988. An integrated approach to project evaluation and selection. *IEEE Transactions on Engineering Management*, 35(4), 265–270.

Khorramshahgol, R., and V.S. Moustakis. 1988. Delphic hierarchy process (DHP): A methodology for priority setting derived from the Delphi method and analytic hierarchy process. *European Journal of Operational Research,* 37, 347–354.

Kline, P. 2000. *An Easy Guide to Factor Analysis*. Routledge, New York.

Koong, K.S., and L.C. Liu. 1999. A factor analysis of attributes affecting computing and information technology usage for decision making in small business. *International Journal of Computer Applications in Technology*, 12(2), 81–89.

Liberatore, M.J. 1987. An extension of the analytic hierarchy process for industrial R&D project selection and resource allocation. *IEEE Transactions on Engineering Management*, EM-34(1), 12–18.

Madu, C.N. 1988. An economic decision model for technology transfer. *Engineering Management International*, 5(1), 53–62.

Masuda, T. 1990. Hierarchical sensitivity analysis of priority used in analytic hierarchy process. *International Journal of Systems Science*, 21(2), 415–427.

Mustafa, M.A. 1989. An integrated hierarchical programming approach for industrial planning. *Computers and Industrial Engineering*, 16(4), 525–534.

Park, H.S., R. Dailey, and D. Lemus. 2002. The use of exploratory factor analysis and principal components analysis in communication research. *Human Communication Research*, 28(4), 562–577.

Rahman, S., and L.C. Frair. 1984. A hierarchical approach to electric utility planning. *Energy Research*, 8, 185–196.

Ramanujam, V., and T.L. Saaty. 1981. Technological choice in the less developed countries: An analytic hierarchy approach. *Technological Forecasting and Social Change*, 19, 81–98.

Russell, D.W. 2002. In search of underlying dimensions: The use (and abuse) of factor analysis in personality and social psychology bulletin. *Personality & Social Psychology Bulletin*, 28(12), 1629–1646.

Saaty, T.L. 1977. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15, 235–281.

Saaty, T.L. 1980. *The Analytic Hierarchy Process*. McGraw-Hill, New York.

Saaty, T.L. 1986. Axiomatic foundations of the analytic hierarchy process. *Management Science*, 32(7), 841–855.

Saaty, T.L. 1990. An exposition of the AHP in reply to the paper "Remarks on the analytic hierarchy process". *Management Science*, 36(3), 259–268.

Saaty, T.L., and H. Gholamnezhad. 1981. Oil prices: 1985 and 1990. *Energy Systems and Policy*, 5(4), 303–318.

Saaty, T.L., and J.M. Alexander. 1989. *Conflict Resolution: The Analytic Hierarchy Approach*. Praeger, New York, 1989.

Saaty, T.L., and L.G. Vargas. 1979. Estimating technological coefficients by the analytic hierarchy process. *Socio-Economic Planning Science*, 13, 333–336.

Saaty, T.L., and L.G. Vargas. 1984. Inconsistency and rank preservation. *Journal of Mathematical Psychology*, 18, 205–214.

Saaty, T.L., and L.G. Vargas. 1987. The analytic hierarchy process: Theoretic developments and some applications. *Math Modelling*, 9, 3–5.

Saaty, T.L., and L.G. Vargas. 1991. *Prediction, Projection and Forecasting*. Kluwer Academic Publishers, Boston, MA.

Saaty, T.L., L.G. Vargas, and R. Wendell. 1983. Assessing attribute weights by ratios. *Omega*, 11(1), 9–13.

Shapiro, S.E., M.R. Lasarev, and L. McCauley. 2002. Factor analysis of Gulf War illness: What does it add to our understanding of possible health effects of deployment? *American Journal of Epidemiology*, 156(6), 578–585.

Spearman, C. 1904. General intelligence: Objectively determined and measured. *American Journal of Psychology*, 15, 201–292.

Snook, S.C., and R.L. Gorsuch. 1989. Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin*, 106(1), 148–154.

Vargas, L.G. 1983. Prediction and the analytic hierarchy process. *Mathematics and Computers in Simulation*, 25, 156–167.

Vargas, L.G. 1990. An overview of the analytic hierarchy process and its applications. *European Journal of Operational Research,* 48, 2–8.

Wabalickis, R.D. 1988. Justification of FMS with the analytic hierarchy process. *Journal of Manufacturing Systems*, 7(3), 175–182.

Wind, Y., and T.L. Saaty. 1980. Marketing applications of the analytic hierarchy process. *Management Science*, 26(7), 641–658.

Winkler, R.L. 1990. Decision modeling and rational choice: AHP and utility theory. *Management Science*, 36(3), 247–248.

Zahedi, F. 1986. The analytic hierarchy process: A survey of the method and its applications. *Interfaces*, 16, 96–108.

VI

Human Factors and Ergonomics

25

Human Factors in Military Systems

Misty Blue
Wright State University

Raymond R. Hill
Air Force Institute of Technology

25.1 Introduction

Many military systems are extremely complex in physical design and operative characteristics. They can utilize a number of people and technologies to complete different types of tasks or missions. For example, military logistics is the science of the planning and moving of military resources. This typically means teams of people coordinating to account for various resources (whether they are in the form of physical equipment and supplies or actual personnel) and determining the most effective use of those resources. Historically, there have been many attempts to make this process more efficient, including just-in-time approaches and enterprise resource planning (ERP) systems. Sense and respond logistics (SRL) incorporates current concepts that rely heavily on automated technologies to provide "adaptive responsive demand and support networks that operate in alternate structures that recognize operational context and coordination" (Office of Force Transformation 2003). This is just one military system that makes use of the interaction between modern technology and the human being. Any time that these two interact, there will be a need for human factors engineering. It is of paramount importance that there is a harmony between the human and the technologies that they use. Human factors engineering, also known as human machine systems engineering, is the most common instrument to develop this necessary harmony.

There is much supporting the need for human factors considerations in even the simplest systems (Darnell 2006; Norman 2002). The objective of this chapter is to describe both human factors engineering and human computer interaction (HCI) and discuss their current applications in the military domain. Specifically, applications to various communication systems, assistive technologies and uninhabited aerial vehicles (UAVs) are discussed. Before human factors in military systems are discussed, a brief description of the military operating environment is given to provide a context for the chapter.

Military systems are very complex in nature. Many aspects of the combat environment are also highly dynamic which adds to its overall complexity. The physical conditions of the environment can range from moderate to extreme in terms of weather, danger, activity, etc. In addition, the number of tasks and task difficulty varies from mission to mission. With so much variation in the working environment, the technologies that are employed during combat must be highly versatile and human factors principles and guidelines must be applied to ensure the safety and effectiveness of the system. This is best achieved when human factors engineering is considered at the beginning stages of product or system development. This ensures that the systems and artifacts implemented will accommodate the human users, and thus enhance the total system performance.

The Natick Soldier Center introduced the approach of soldier-centric "system of systems". This concept is centered on the integration of the human and all soldier-borne systems (Brandler, 2005). This approach considers the soldier a system in itself. Brandler (2005) states "…from a broad perspective this soldier-oriented system of systems must be compatible and interface with large, complex systems with which the soldier interacts on the battlefield such as helicopters, tanks, and other vehicles and most recently, with the "network". The soldier system and all it comprises must work synergistically with those large systems." When these types of user-centered principles are utilized in the research and development phases of military programs and operations, the returns on the investment are found in the safety of our troops, enhanced soldier performance, and effective systems.

The US military is moving towards a network centric environment that is made possible by a wide range of technologies designed to assist and enhance the overall performance of the soldier and all combat equipment. All assistive technologies must in some way enhance the performance and/or survivability of the soldier. They must be able to accommodate a broad range of environments and people to be effective. Various aspects of human factors research define this range of environments and people along with the limitations of each. As this information is acquired, we can progress towards optimal human–machine systems.

The following section provides a brief history and background of human factors and human factors engineering.

25.2 Human Factors: Background and History

A widely accepted definition of human factors is presented by Sanders and McCormick (1993):

> *Human factors discovers and applies information about human behavior, abilities, limitations, and other characteristics to the design of tools, machines, systems, tasks, jobs, and environments for productive, safe, comfortable, and effective human use.*

The procedures and techniques involved in human factors analyses involve two major components: the human and the system or object being analyzed. The major goals of these analyses are to create tools and tasks that increase the effectiveness and efficiency of the interaction between the human and the system, thus increasing the quality and productivity of work produced by the system. In addition, the analyses should aim to improve the working conditions for the human. By this, we mean that the working environment is safe, user friendly, not cognitively or physically stressful, and relatively comfortable. Many human factors and ergonomics researchers have spent their careers developing a body of knowledge that defines and represents the human limitations, capabilities, and behaviors. This body of knowledge is used as fundamental knowledge in the design of systems, both civilian and military. This body of knowledge attempts to fit the artifacts and tasks of the system to the human, and not the other way around; this is the basic idea behind user-centered design.

The history of human factors dates back to the early 1900s with the work of Frank and Lillian Gilbreth with it becoming a formal profession in the mid and late 1940s. In 1949, the first book in the field was published, entitled *Applied Experimental Psychology: Human Factors in Engineering Design* (Chapanis et al. 1949). Not surprisingly, given the nature of most military operations, most of the growth in the field of human factors for the next 20 years was in the arena of military systems research. In a sense, human factors has grown in many different directions and is now more of an umbrella discipline that covers a wide spectrum of related disciplines. Some of these disciplines include cognitive engineering, workload analysis, decision making, displays, ergonomics, and operations engineering.

In the past, human factors analysts have had to justify their roles in research and development projects (Price et al. 1980). Currently, human factors research is a major part of many military endeavors. All branches of the US military now have at least one human factors branch. In addition, the US Department of Defense has organized the Human Factors Engineering Technical Advisory Group (HFE-TAG) to identify and address human factors issues in both the military and civilian domains. The following section will discuss some of the efforts in terms of human factors that the US military is currently embarking on. This is intended to be an overview of just some of the human factors issues being addressed by the US military.

25.3 Human Factors and Communication Systems and Other Assistive Technologies

25.3.1 Military Communication Systems

Advancements in technology have revolutionized military communication systems. These systems function as vehicles for mutual information exchange between humans, between machines, and between the human and the machine. Computer technology allows communication systems to link radio lines with computers to format information, distribute it within headquarters and record it electronically (Ripley, 1997). With advances in speech technologies, a number of new devices are of interest to armed forces. Some of the current communication systems and their related human factors issues are discussed in the following paragraphs.

Vast amounts of military communications depend on real time data from UAVs, various military assets, and soldiers on the ground. Operations in the global war on terror have demonstrated the major bottleneck associated with communication bandwidth. The problem is mostly technical but also creates major human factors issues. When the connection between soldiers and their means of communications is not effective, there is an increased risk of loss of situation awareness and the inability to direct and receive directions for ground soldiers. Situation awareness is the perception and comprehension of elements in the environment and the prediction of their future status (Endsley and Garland 2000). There is danger to soldiers when their situation awareness is impaired in any way. To address this issue the US Department of Defense formed the Transformational Communications Architecture (TCA). A major part of the effort is the Transformational Communications Satellite System (TSAT). It was reported that "TSAT is intended to provide internet-like capability that extends high-bandwidth satellite capabilities to deployed troops worldwide, and delivers an order of magnitude increase in available military bandwidth. Using laser communications intersatellite links to create a high data-rate backbone in space, TSAT will be one of the key enablers for the American vision of Network Centric Warfare" (Defense Industry Daily 2007). Constant and accurate information is a major key in maintaining situation awareness. Satellite technologies help enhance communications between military personnel.

25.3.2 Future Force Warrior

The US Army is focusing their attention on the concept of the future force warrior (FFW). The impetus for the FFW is to provide a ten-fold increase in both lethality and survivability of the infantry platoon by enhancing situational awareness, incorporating precise and effective firepower and netted communications (Future Force Warrior Infantry Combat Suite 2004). Brandler (2005) states that the "primary principles of the FFW program are to accomplish the following using a human-centric design paradigm:

- Provide collaborative capabilities distributed across the small combat unit.
- Reduce the weight of the individual warfighter and the small combat unit.
- Amplify and extend the ability to see and hear first.

- Rapidly know, understand, decide and adjust on the move.
- Multiply the ability to be overwhelmingly lethal.
- Extend the ability to survive and endure longer missions.
- Optimize the concept of operations for the small combat unit—*Change the way we fight.*

The FFW approach is an integrated system of systems approach in support of a soldier-centric force. The FFW combat uniforms integrate various technologies that provide both protection to soldiers during battle along with advanced means of communications. Figure 25.1a shows Future Force Warrior Advanced Technology Demonstration (FFW ATD). This ensemble is fully integrated and lightweight. The soldier has full body protection, netted communications, a wearable power source, and a weapon system. With the ensembles' capabilities distributed across small teams, the performance of the small combat team can be greatly improved. According to the US Army Natick Solider Research, Development, and Engineering Center (2007), the FFW ATD is realized in the capabilities of lethality (software and hardware for fire control of integrated lightweight weapons), survivability (multi-functional, full spectrum protective ensemble with physiological/medical sensor suite and laser eye protection), sensors and communications (advanced team communications), power and mobility. Figure 25.1b demonstrates the Future Warrior Concept (FWC). At this point in time, this is still just a concept that is being used as a springboard to begin considerations of properly equipping ground soldiers in the future. According to the US Army Natick Solider Research, Development and Engineering Center (2007), there are six major subsystems incorporated in the concept: (1) the headgear subsystem, (2) the combat uniform subsystem,

FIGURE 25.1 Future Force Warrior ensembles. (a) Future Force Warrior Advanced Technology (FFW ATD); (b) Future Warrior Concept (FWC). (Source: http://nsrdec.natick.army.mil/media/photo/index.htm.)

(3) the weapon subsystem, (4) the war fighter physiological status monitor subsystem, (5) the micro-climate conditioning subsystem, and (6) the power subsystem.

During the fall of 2006, the FFW ensemble was subjectively evaluated by the Human Research and Engineering Directorate of the US Army Research Laboratory at Fort Benning, GA (Turner and Carstons 2007). The assessment was broken up into three phases: training, pilot exercises, and an air assault expeditionary force experiment. Trained soldiers compared two FFW ensembles (Commander or leader and soldier designs) to the equipment that they normally used using activities within their normal routine. Table 25.1 summarizes the components of the systems evaluated.

TABLE 25.1 Brief Descriptions of Components of the FFW

System Component	Brief Description	Ensemble Variation
Multi-function combat suit (MFCS)	Multi-cam camouflage shirt and pants with integrated knee and elbow pads.	Leader/soldier
Soldier protection integrated ensemble system (SPIES) chassis	Integrated body armor and load carriage system. Holds all test related electronics and hardware. Also housed front and back protective training plates.	Leader/soldier
SPIES ballistic load belt	Used to carry additional mission-critical objects.	Leader/soldier
Hydration system	Enabled soldiers to carry water while moving. Worn to the right of the back plate carrier of SPIES chassis during testing.	Leader/soldier
Push to talk and body-worn antenna	Allows wearer to communicate with squad leader. Worn on the front side of the left shoulder. Body-worn antenna enabled robust narrowband and wideband communications capabilities. Worn on the front protective plate of the SPIES chassis.	Leader/soldier
Global-positioning system (GPS)	Captured satellite data on the wearers coordinates. Worn on the back side of the left shoulder.	Leader/Soldier
Wearable soldier radio terminal (WSRT)	Established communications network connectivity within the FFW squad and to higher elements. Used to send situation awareness, command and control data, and voice communications within the squad. Worn in a pouch behind left arm.	Leader/soldier
Computer	Used to process GPS data and XM-104/multi-functional laser data in the Falcon View[1] software package. Housed in computer carrier on the rear of the chassis.	Leader/soldier
Trackball mouse	Used to manipulate the Bare Bones and Falcon View software that appeared in the goggle-mounted display. Carried in a pouch typically on the right hand side of the chassis.	Leader/soldier
FFW helmet	Provided ballistic and impact protection, and communication and mounting functions. Communications were enhanced with two bone conduction speakers (one on each side of helmet). Boom microphone provided voice communications. Night vision goggles on top of the helmet supplied hands-off enhanced vision during low light conditions.	Leader/soldier
Personal digital assistant	Provided SA to the soldier via government-owned C2 MINCS software. PDA had its own battery attached to the bottom. Stored in a pouch on the left-hand side of chassis under soldiers left arm.	Soldier
Lithium-ion battery	Power WSRT and GPS. Lasted more than 12 hours. Stored in pouch below assault pack.	Soldier
Soldier system headgear	Microphone extended off the FFW helmet. Two bone conduction speakers attached to the supporting strap at the temples allowed incoming voice messages without disrupting ambient hearing.	Soldier
Command and control mobile intelligent network-centric computing system (MINCS) soldier software	Dismounted computing platform designed to provide network-centric C4ISR connectivity. Provided dismounted fighters continuous real-time SA of friendly locations, tactical report generation, capability to communicate with higher ranking personnel, memory joggers, and integration with MFL for target transfer.	Soldier

(*Continued*)

TABLE 25.1 Brief Descriptions of Components of the FFW (Continued)

System Component	Brief Description	Ensemble Variation
Leader computer	Used to process GPS data and XM-104/multi-functional laser data in the Bare Bones and Falcon View software package. Housed in computer carrier on the rear of the chassis.	Leader
Multi-function laser (MFL) (squad leader and rifleman only	Designed to enhance target engagement. Attached to weapon and to the PDA or leader computer to enable aim point and import/export of target information across the network via C2 MINCS.	Leader
XM-104	Fire control system designed to enhance target engagement during cooperative engagements. Enabled aim point and import/export of target information across the network. Connected to the leader computer.	Leader
Rechargeable lithium-ion battery (2)	Power WSRT, GPS, and computer. Second battery provided as a backup. Both batteries lasted approximately 12 hours. One was stored in the computer carrier and the other in a pouch behind soldiers right arm.	Leader
Battlefield renewable integrated tactical energy system (BRITES) power manager	Managed power consumption of the two batteries by the electronics.	Leader
Leader computer	Panasonic CF-18 Toughbook computer; provided all processing and hardware to run Falcon View leader software. Stored in the backpack chassis and connected to the goggle-mounted display and trackball mouse.	Leader
Leader system headgear	Two parallel-mounted microphones (one for hands free communication and the other to issue verbal commands to the Falcon View and Bare Bones software programs) extended of the FFW helmet. Multi-functional color goggle mounted display displayed tactical processor information.	Leader
Falcon View leader software	Government-owned Microsoft Windows[2] based mapping application that displays various types of maps and geographically referenced overlays. Provided dismounted soldiers continuous real-time SA of friendly locations, tactical report generation, communications capabilities with higher echelons, UAV aero-environment controls, system voice control, Bare Bones targeting system, route planner, and memory joggers.	Leader
Trackball mouse	Used to manipulate the Bare Bones and Falcon View software that appears in the goggle mounted display. Carried in a pouch, typically on the right-hand side of the chassis.	Leader

Source: Turner, D. and Carstens, C. 2007. Future Force Warrior: Insights from Air Assault Expeditionary Force Assessment. Army Resea Research Laboratory, Aberdeen Proving Ground, MD. Technical Report No.: ARL-TR-4191. With permission.

[1] Falcon View is a trademark of Georgia Tech Research Institute.

[2] Windows is a trademark of Microsoft Corporation.

Overall, the FFW system was considered a major success. The tested soldiers strongly preferred the FFW equipment to their baseline gear for virtually every activity in every scenario. The test report did note the possibility of cognitive and perceptual overload of the visual and auditory senses. Possible tactile displays may be considered in the future as a result of the recommendations provided. This is a very noteworthy achievement for the Department of Defense considering the improvements from evaluations of previous versions (Turner et al. 2005) and of the disappointing evaluations of the Land Warrior system (Erwin 2006).

When these systems are in use, the individual soldiers will have access to network-centric information. One way this is made possible is through rugged personal digital assistants (R-PDA's) designed for maintaining situational awareness and conducting mission planning. These can have a range of functions including the Blue Force Tracking Technology that helps to identify friendly forces, satellite phone capabilities, and downloadable maps with the ability to download instructions overlaid on the maps. Both situational awareness and mission planning are aided with various displays of icons and terrain or navigational maps. When displays of dynamic systems are employed, the displays must be easy to interpret and assist the user in knowing, understanding, and predicting future events and conditions, all of which are of paramount importance to military operations. Figure 25.2 shows examples of PDAs being produced for military operations. Human factors evaluations of these devices have unveiled various issues with the display and the difficulty in manipulating the piece of equipment while in transit (Turner and Carstons 2007). Since it is a hand held device, the militarized PDA's have a small display screen that needs to provide a lot of information.

A major branch of human factors focuses on visual displays. The objective of any visual display is to support the perception of relevant data and facilitate the further processing of that information (Wickens et al. 2004). As displays get smaller, human factors analysts must re-evaluate guidelines of visual displays. Research has shown that when small screens are used for retrieval tasks, people were 50% less effective in completing the tasks than those using larger screens (Jones et al. 2007). In terms of navigational displays, Garland and Endsley (1995) states that the display should support (1) guidance on how to get to a destination, (2) planning facilitation, (3) help with recovery when user is lost, and (4) maintain situation awareness regarding the location of a broad range of objects. The CDA certainly meets these basic objectives, but there are a number of recommendations from the literature for the improvement of the human factors of the technology. Turner and Carstons (2007) offer the following potential solutions.

- When icon sizes are too small on the screen, a zoom in feature or a text description when the cursor moves over the icon could be used to increase visibility.
- Simplify the iconography used in the display.
- Increase refresh rate of displays.

The concept of the PDA has also been evaluated as a training aid. The US Navy performed a usability evaluation on a PDA for the shipboard training environment (Allen et al. 2003). The study revealed several limitations when PDAs are used for US Naval data collection efforts ranging from visual display issues (mostly related to the screen size) and input methods. As technology advances, there is a great potential for the PDA in a variety of military applications. But this potential will only be realized once human factors issues related to visual display effectiveness are resolved.

(a) (b)

FIGURE 25.2 Examples of Commander's Digital Assistants (CDA). (a) US Army PDA (Photo taken from http://www.strategypage.com/military_photos/2004620.aspx); and (b) Commander's Digital Assistant (CDA) produced by General Dynamics.

25.4 Human Factors and Unmanned Aerial Vehicles (UAVs)

UAVs, also known as remotely piloted aircraft (RPAs) or unmanned aircraft systems (UASs), are remotely piloted (or in some cases self-piloted) aircraft that can carry and use various resources (ammunition, cameras, sensors, etc.), can be used for surveillance purposes, or can be used for offensive operations. Many missions that are otherwise too dangerous or impossible to deploy with manned aircraft are possible with UAVs (Gawron 1998). The US Department of Defense defines them in the following manner (Bone and Bolkcam 2003).

> *Powered, aerial vehicles that do not carry a human operator, use aerodynamic forces to provide vehicle lift, can fly autonomously or be piloted remotely, can be expendable or recoverable, and can carry a lethal or nonlethal payload.*

UAVs have tremendous potential for the military in that they provide a means to remove infantry from harms way. Also, some have capabilities of longer flights and can sustain higher G-forces (since manned aircraft are G-limited by the on-board human). The Department of Defense (2005) reported that as of September 2004, 20 types of coalition uninhabited aircraft, large and small, have flown over 100,000 total flight hours in support of Operation Enduring Freedom (OEF) and Operation Iraqi Freedom (OIF). Table 25.2 shows the various categories of UAVs currently in production, listed by the UAV forum (SRA International, 2006).

There are a number of human factors issues associated with the employment of UAVs. Their operation represents a highly complex automated system whose success depends on cooperation between the human operator, the physical UAV, and the software/hardware controller systems. Some of the traditional human factors research is based on manned vehicles and cockpits and is not readily applicable to unmanned vehicles. Thus, there is a need for research focusing on the issues of unmanned vehicles. Table 25.3 as reported by Tvaryanas (2006), shows a comparison of the human factors concerns for manned versus unmanned vehicles from the perspective of aerospace medicine human factors. Many of the factors are relevant for both the manned and unmanned systems but the context is obviously different. The historical principles from the manned cockpit are merely baseline when it comes to UAVs. Since there is a need for more research in the human factors area of UAVs, the following sections discuss some of the current research in progress.

The issues discussed in the following paragraphs include those related to automation and control, human-machine interface (HMI), and mishaps.

TABLE 25.2 Categories of UAVs in Production

Classification	Brief Description
Tactile	
Endurance	Capable of flights greater than 24 hours.
Vertical takeoff and landing	Typically uses a rotary wing to takeoff and land vertically.
Mini also known as man portable	Larger than micro air vehicles but small enough to be carried by one person in a backpack; Usually launched by hand throwing or sling shot instrument.
Optionally piloted vehicle	Can be operated manned or unmanned; usually adapted general aviation aircraft.
Micro air vehicle	Has no dimensions larger than 6 inches.
Research	Developed for specific investigations, typically not intended for production.

TABLE 25.3 Comparison of Aerospace Medicine Human Factors Concerns for Manned Aircraft (MA) versus Remotely Piloted Aircraft (RPA) Crewmember Performance

Factors	MA	RPA*	Factors	MA	RPA
Physical environment	+	+	Motion sickness	+	±
Vision restricted (clouds, ice, etc.)	+	+	Hypoxia and hypobarics	+	0
Noise and vibration	+	±	Visual adaptation	+	±
Windblast	+	0	Physical task oversaturation	+	+
Thermal stress	+	±		+	+
Maneuvering forces	+	0	Perceptual factors	+	0
			Illusion-kinesthetic	+	0
Technological environment	+	+	Illusion-vestibular	+	+
Seating and restraints	+	0	Illusion-visual	+	+
Instrumentation	+	+	Misperception of operational conditions	+	+
Visibility restrictions (e.g. FOV)	±	+	Misinterpreted/misread instrument	+	+
Controls and switches	+	+	Spatial disorientation	+	+
Automation	+	+	Temporal disorientation	+	+
Personal equipment	+	0			
Cognitive	+	+			
Vigilance and attention management	+	+	Crew coordination and communication	+	+
Cognitive task oversaturation	+	+	Distributed / virtual crew	0	±
Confusion	+	+	Shift changeovers	0	±
Negative transfer	+	+			
Distraction	+	+			
Geographic misorientation (lost)	+	+	Self-imposed stress	+	+
Checklist interference	+	+	Physical fitness	+	+
			Alcohol	+	+
Psycho-behavioral	+	+	Drugs, supplements, or self medications	+	+
Personality style	+	+	Inadequate rest	+	+
Emotional state	+	+	Unreported disqualifying medical condition	+	+
Overconfidence	+	+			
Complacency	+	+			
Motivation	+	+			
Burnout	+	+	Miscellaneous		
			Multi-aircraft control	0	±
Adverse physiological states	+	+	Control and feedback latency	0	+
Effects of G-forces	±	0	Standardized cockpit design and controls	+	0
Prescribed drugs	+	+	Manual control of aircraft	+	±
Sudden incapacitation	+	+	Standardized crew qualifications	+	0
Pre-existing illness or injury	+	+	"Shared fate" with aircraft	+	0
Physical fatigue	+	+			
Mental fatigue	+	+			
Circadian desynchrony	+	+			

Source: Tvaryanas, A. 2006. Human systems integration in remotely piloted aircraft operations. *Aviation Space and Environmental Medicine*, 77, 1278–1282. With permission.

+ = usually applicable; ± = possibly applicable; 0 = not applicable.

* If an RPA is operated from another airborne platform, all MA performance concerns would also apply.

25.4.1 UAV Automation and Control

Flight control of UAVs is a highly automated procedure. The aircrew and aircraft are not in the same location. Flight control and system performance maintenance are made possible through various levels of automation. As with most systems, there are both strengths and shortcomings to automation. According to Hopcroft et al. (2006), "optimal flight performance will require that the operators be aware of and understand the activities of the automated system, and that the actions of operators and the automated system complement rather than compete with one another." Automation allows for certain functions to be performed by a machine (Parasuraman and Riley 1997). Having some functions automated can release the human operator from such tasks and permit them to reallocate cognitive and physical resources to execute other tasks and possibly reduce the potential of operator overload (Sarter et al., 1997). However, when implementing automation in UAV control, careful consideration of the design of the system and mission requirements should be taken into account as inadequately planned automation can actually increase operator workload and decrease situation awareness (Ruff et al. 2002).

When a system is highly automated, dependence and trust issues can come into play. Dependency here is an over-reliance on the automation. Trust has been defined in a variety of contexts but in general is the extent to which people believe the trusted object or person is dependable. There is a wealth of research available on trust in automation in a number of domains. The core of the literature is typically credited to Parasuraman and Riley (1997) with their discussion of automation use, misuse and disuse. Trust represents the confidence the operator has in the technology being used which has an affect on their attitude about using it. With this, trust is more of an issue of how the human operator feels about the system in use and this trust determines whether or not they will use the automated features or not. Trust issues tend to apply more to systems with lower levels of automation (the operator has more options when it comes to automation, i.e. more manual features are available). If operators mistrust automation, they may not take advantage of it and depend more on their own abilities. If operators have too much trust in the system, they may over-rely on the automatic features.

System transparency is achieved when the user can view the interface of the system and quickly get an understanding of system state (what the system is doing and why) (Mark and Kobsa 2005). On the other end of the automation spectrum, higher levels of automation (system makes most decisions without aid from the human operator) tend to decrease system transparency, which can lead to a decrease in situation awareness (Hopcroft et al. 2006). The balance between the amount of automated features and the amount of manual is a very delicate that still needs more research specifically in the areas of which tasks are most appropriate to automate and to how much automation is suitable (McCarley and Wickens 2005).

25.4.2 UAV Human Machine Interface

The HMI is the equipment that the operator interacts with to control the UAV. The HCI is a branch of human factors concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them (ACM SIGCHI 1999). HCI includes several areas that are relevant to the successful use of UAVs, mostly related to the displays utilized by the ground control centers. Wilson (2002) states "in designing an optimal HMI for UAV and UCAV operations, an essential issue is information display. Options range from conventional 2-D screens to flat perspective-view visualizations to multi-sensory 3-D immersive displays promoting virtual presence." Figure 25.3 shows three different control centers from General Atomics Aeronautical Systems developers of the Predator UAV system. The actual control center can have a range of configurations from somewhat complicated to very complicated. Control centers such as those depicted in Figure 25.3a and b can introduce problems such as cognitive overload with the use of multiple screens, cognitive tunneling, and fatigue. Even the more simplified configurations (Figure 25.3c) can have HCI, ergonomic,

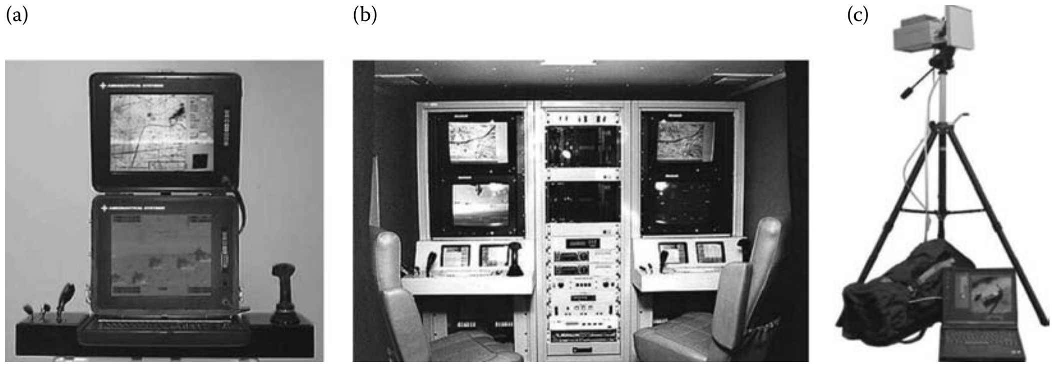


FIGURE 25.3 Ground control centers from General Atomics Aeronautical Systems. (a) Portable Ground Control Station (PGCS), (b) ground control system for the US Predator Military Unmanned Air Vehicle, and (c) Portable Ground Control Station. (Source: http://www.ga-asi.com/resources/image_library/ground_control.php.)

and human factors implications that if not addressed can negatively impact mission success. There are a number of guidelines and principles that are available in the literature for information display and display design. These are summarized in the following paragraphs.

Wickens et al. (2004) list 13 principles of display design in four different categories: (1) perceptual principles, (2) mental model principles, (3) human attention principles, and (4) human memory principles. The perceptual principles are as follows:

1. *Make displays legible*: For a display to be useful, it must be legible in terms of contrast, visual angle, brightness, illumination, noise, masking, etc.;
2. *Avoid absolute judgment limits*: Lists longer than five or seven items are not easily retained without some organizational structure. Do not require operators to make judgments on the levels of a variable in the display that has more than five or six levels.
3. *Top-down processing*: To guarantee that a signal is interpreted correctly, it must be presented in accordance with operators expectations regarding that signal.
4. *Redundancy gain*: In degraded communications environments, messages that are repeated at least once are more likely to be interpreted correctly (particularly if presented in an alternative physical form).
5. *Discriminabilty*. Similarity causes confusion.

The next set of principles are those associated with the mental model:

6. *Principle of pictorial realism* (Roscoe 1968): The display should look like the variable it represents.
7. *Principle of the moving part* (Roscoe 1968): Any moving elements of a display should move in accordance with the operator's mental model of how the variable moves in the physical system.

The principles that are based on attention are:

8. *Minimizing information access cost*: Minimize the cost of time and effort to move selective attention from one display to another by keeping frequently used display elements located where the cost of traveling between them is small.
9. *Proximity compatibility principle*: If two or more sources of information need to be mentally integrated to accomplish a task, the sources of information should be in "close proximity" of each other. This can be achieved with space, color, lines linking the sources of information together, etc.
10. *Principle of multiple resources*: When a lot of information has to be processed, dividing it across resources (visual, auditory, tactile) can be helpful.

The final three principles are those related to human memory:

11. *Replace memory with visual information: knowledge in the world*: Do not require operators to retain important information solely in their working memory or retrieve it from their long-term memory.
12. *Principle of predictive aiding*: Displays that can clearly predict what will happen or what is likely to happen are very effective in enhancing human performance.
13. *Principle of consistency*: Design displays in a manner that is consistent with other displays the operator is viewing concurrently.

In addition to the guidelines previously described, the Federal Aviation Administration lists a number of guidelines and criteria for displays (Ahlstrom and Kudrick 2007). The following lists those most relevant to UAVs. There are a number of others that are useful in the design of control center displays. For a comprehensive list of guidelines, refer to the appendix of Ahlstrom and Kudrick (2007).

1. *Make displays function under operational conditions*: Visual displays shall function under any circumstance corresponding with the operational and use philosophies of the system.
2. *Evaluate through prototyping*: The suitability and effectiveness of a display should be evaluated using representative tasks, users, and environmental settings before being incorporated in a new system.
3. *Reliability*: The mean time between failure goal of the display should be more than 10,000 hours.
4. *Aspect ratio*: The aspect ratio of a display is the ratio between the horizontal and vertical dimensions of the display. This ratio should not adversely impact displayed data. Common ratios include 4:3, 4:5, and 16:10.
5. *Adjust to user comfort*: Users should be able to easily angle, tilt, and swivel a display to maintain a comfortable working position.
6. *Simultaneous use*: A visual display that must be monitored concurrently with manipulation of a related control shall be located so that it can be read to within required accuracy while adjusting the control.
7. *Multiple displays*: When the manipulation of one control requires the reading of several displays, the control shall be placed as near as possible to the related displays, but not so as to obscure displays when manipulating the control.
8. *Arrange according to function and sequence*: Displays shall be arranged in relation to one another according to their sequence of use or the functional relations of the components they represent.
9. *Determine the maximum viewing distance by the legibility*: Maximum viewing distance for displays should be determined by the legibility of the displayed information.
10. *Detection of weakest target*: Adjustment of brightness, contrast, and other electronic parameters shall permit the detection of the weakest target that is simulated.

By definition, the operator and the physical UAV are not colocated. This calls for more advanced human factors guidelines for UAVs. The guidelines borrowed from manned aviation are an excellent starting point, but for UAV operations to be as successful as manned aircraft, an advanced interface that provides an accurate and high level of situation awareness to the operator while avoiding any physical and cognitive overload is going to be a necessity. Efforts at the Air Force Research Laboratory are focused on developing human-centered technologies and providing human factors design criteria for the development of the HMI. General descriptions of research activities can be found on the web (http://www.wpafb.af.mil/shared/media/document/AFD-070418-027.pdf). To summarize of the activities of the research group, the systems control interfaces branch (RHCI) has three main programs; Synthetic Interfaces for Unmanned Air Systems Research (SIRUS), Vigilant Spirit Control Station (VSCS), and Flight Display Integration for Autonomous Approach and Landing Capability (AALC). The (SIRUS) program is currently evaluating the incorporation of immersive and 3-D perspective technologies in the interface. SIRUS is also developing and evaluating interactive map displays, perspective views and immersive interface concepts that support future upgrades to the Predator system of UAVs. The VSCS program is evaluating operator–vehicle interface technologies required to control, manage, and operate multiple UAVs with minimal crew size. The AALC is working on advanced symbology for heads up displays.

25.4.3 UAV Mishaps

As with any system, UAV safety is of utmost importance. Even though these systems take the human being out of danger, these aircraft are expensive to build and repair. Thus, the safety and reliability of the aircraft itself has become an important research area. As the use of UAVs increases, the high rate of mishaps has become more obvious. The Defense Science Board of the Office of the Secretary of Defense (2004) reported high mishap rates as one of the two greatest reasons the military may not realize the full potential of remotely piloted aircraft (unit cost was the second reason). When comparing with traditional manned vehicle mishaps (one mishap per 100,000 flight hours) with UAV mishaps (32–344 mishaps per 100,000 flight hours depending upon the vehicle) the disparity in the reliability of the aircraft types becomes evident. As previously stated, there are a number of UAVs being employed and the human factors issued and concerns vary amongst the different models (Williams, n.d.).

Tvaryananas et al. (2006) performed an epidemiological study using a standardized human factors taxonomy and hierarchical model of human error from the Department of Defense to define a distribution and the determinants of operator error in RPA mishaps within the US Air Force, Army, and Navy/Marines. A 10-year (1994–2003) cross sectional quantitative analysis was performed. Their findings showed "recurring human factors failure at the organizational, supervision, preconditions, and operator levels have contributed to more than half of RPA mishaps." The most pronounced latent failures were observed at the organizational levels. As defined in the Department of Defense Human Factors Analysis and Classification (HFACS) tool, organizational influences on mishaps are "factors in a mishap if the communications, actions, omissions or policies of upper-level management directly or indirectly affect supervisory practices, conditions or actions of the operator(s) and result in system failure, human error or an unsafe situation." These failures were related to operator error and mechanical failures. The results from this study lend to the need for more research on the development of better procedural knowledge for those in management positions of UAV control. This study was a cross sectional study across the various branches of the US military. The results were in line with several previous studies mentioned in the report. Table 25.4 provides a summary of some prior studies from the individual branches of the US military. All studies show a rate of human factors related errors of at least 32%. In addition, the studies that utilized the Taxonomy of Unsafe Acts (Seagle 1997; Ferguson 1999) reported relatively high rates of unsafe supervision which can correlate to organizational influences in the HFACS taxonomy.

25.5 Summary

There are many human factors considerations that should be accounted for during the development and implementation of military systems. Since the nature of the military system is extremely complex and dynamic, human factors engineering is necessary to effectively integrate the intricate components of the system. One method of achieving user-centered design is to view the soldier as a system. It is essential to consider the human as apart of the system during all stages of development rather than designing systems and artifacts and assessing their performance on users as a final evaluation.

The principles of human factors when properly applied can enhance soldier performance and increase system safety and survivability. Some of those considerations were discussed here as related to some military communications systems, the FFW, and UAVs with the common thread being increasing situation awareness. The goal of the FFW is a ten-fold increase in lethality and survivability through the use of an integrated system of systems approach. There are human factors matters ranging from the fit

TABLE 25.4 Summary of Prior Rpa Mishap Studies

Seagle (1997)	Ferguson (1999)	Manning et al. (2004)	Rogers et al. (2004)
Navy $n=203$	Navy $n=93$	Army $n=56$	Air Force, Army $n=48$
Taxonomy: Taxonomy of unsafe acts	**Taxonomy:** Taxonomy of unsafe acts	**Taxonomy:** HFACS	**Taxonomy:** Human systems issues
Human factors: 43%	**Human factors:** 59%	**Human factors:** 32%	**Human factors:** 69%
Factors	*Factors*	*Factors*	*Factors*
Unsafe acts (59%)	Unsafe acts (38%)	Unsafe acts (61%)	Training (27%)
Accidental acts (52%)	Intended (17%)	Skill based (22%)	Team performance (25%)
Slips (2%)	Mistakes (12%)	Decision (33%)	Situational awareness (68%)
Lapses (16%)	Violations (7%)	Misperception (17%)	Interface design (16%)
Mistakes (39%)	Unintended (20%)	Violations (11%)	Cognitive and decision making (14%)
Conscious acts (7%)	Slips (14%)	Preconditions (6%)	
Infractions (6%)	Lapses (3%)	CRM (6%)	
Unsafe conditions (46%)	Unsafe conditions (40%)	Unsafe supervision (50%)	
Aeromedical (20%)	Aeromedical (10%)	Inadequate supervision (33%)	
CRM (27%)	CRM (28%)	Failed to correct known problem (17%)	
Readiness violations (7%)	Readiness violations (10%)	Supervisory violations (11%)	
Unsafe supervision (61%)	Unsafe supervision (43%)	Organizational influences (44%)	
Unforeseen (34%)	Unforeseen (15%)	Organizational processes (44%)	
Foreseen (47%)	Foreseen (12%)		

Source: Adapted from Tvaryanas, A., Thompson, W., and Constable, S. 2006. Human factors in remotely piloted aircraft operations: HFACS analysis of 221 mishaps over 10 years. *Aviation Space and Environmental Medicine*, 77(7), 724–732.

of the ensemble to the operations of the various aspects of the system. In addition, the human factors considerations of UAVs in terms of automation and control, HMI, and mishaps were discussed.

References

ACM SIGCHI. 1999. ACM SIGCHI Curricula for Human Computer Interaction. Retrieved December 15, 2007 from http://sigchi.org/cdg/cdg2.html#2_1.

Ahlstrom, V. and Kudrick, B. 2007. Human factors criteria for displays: A human factors design standard update of chapter 5. Federal Aviation Administration, Atlantic City, NJ. Technical Report No. DOT/FAA/TC-07/11.

Allen, R., Hession, P., and Kring, E. 2003. Usability analysis of a personal digital assistant based data collection tool for the shipboard training environment. Proceedings of RTO HFM Symposium on Advanced Technologies for Military Training. Genoa, Italy.

Bone, E. and Bolkcam, C. 2003. Unmanned Aerial Vehicles: Background and Issues for Congress. Retrieved November 8, 2007 from http://www.fas.org/irp/crs/RL31872.pdf.

Brandler, P. 2005. The United States Army future force warrior—an integrated human centric system. *Strategies to Maintain Combat Readiness during Extended Deployments—A Human Systems Approach*. (pp. KN1–KN12). Proceeding RTO-MP-HFM-124. Prague Czech, Republic on October 3–5, 2005.

Chapanis, A.R.E., Garner, W.R., and Morgan, C.T. 1949. *Applied Experimental Psychology: Human Factors in Engineering Design.* Wiley, New York, NY.

Darnell, M. 2006. Bad Human Factors Designs. Retrieved on November 8, 2007, from http://www.baddesigns.com/.

Defense Industry Daily. 2007. Lockheed gets Another $41.7M for TSAT. Retrieved on November 8, 2007 from http://www.defenseindustrydaily.com/lockheed-gets-another-417m-for-tsat-0351/.

Department of Defense. 2005. Unmanned Aircraft Systems Roadmap 2005–2030. Retrieved November 8, 2007 from http://www.fas.org/irp/program/collect/uav_roadmap2005.pdf.

Department of Defense Human Factors Analysis and Classification System. 2007. A Mishap Investigation and Data Analysis Tool. Retrieved December 6, 2007 from http://www.uscg.mil/hq/g-w/g-wk/wks/pdf_files/hfacs.pdf.

Erwin, S. 2006. Army Land Warrior not ready for war, gears up for next tryout. National Defense. Retrieved December 22, 2007 from http://goliath.ecnext.com/coms2/gi_0199-5318709/Army-Land-Warrior-not-yet.html#abstract.

Endsley, M. and Garland, D. 2000. *Situation Awareness Analysis and Measurement*. Lawrence Earlbaum Associates, New York, NY.

Ferguson, M. 1999. Stochastic modeling of naval unmanned aerial vehicle mishaps: Assessment of potential intervention strategies [Thesis]. Naval Postgraduate School, Monterey, CA.

Future Force Warrior Infantry Combat Suite. 2004. Defense Update. Issue 5.

Garland, D., Endsley, M. (Eds.). 1995. Proceedings of an International Conference: *Excremental Analysis and Measurement of Situation Awareness.* Embry Riddle Aeronautical University, Daytona Beach, FL.

Gawron, J. 1998. Human factors issues in the development, evaluation and operation of uninhabited aerial vehicles. Proceedings of the Association for Unmanned Vehicle Systems International. Arlington, VA.

Hopcroft, R., Burchat, E., and Vince, J. 2006. *Unmanned Aerial Vehicles for Maritime Patrol: Human Factors Issues*. DSTO Defense Science and Technology Organization, Commonwealth of Australia.

Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K., and Buchanan, G. 2007. Improving Web Interaction on Small Displays. Retrieved November 12, 2007 from http://www8.org/w8-papers/1b-multimedia/improving/improving.html.

Manning, S., Rash, C., LeDuc, P., Noback. R., and McKeon, J. 2004. The role of human causal factors in U.S. Army unmanned aerial vehicle accidents. US Army Aeromedical Research Laboratory, Ft. Rucker, AL. Report No.: USAARL-2004-11.

Mark, G. and Kobsa, A. 2005. The effects of collaboration and system transparency on CIVE usage: An empirical study and model. *Presence: Teleoperators and Virtual Environments,* MIT Press 14(1), 60–80.

McCarley, J. and Wickens, C. 2005. Human factors implications of UAVs in the national airspace. Aviation Human Factors Division, Technical Report No. AHFD-05-05/FAA-05-01.

Norman, D. 2002. *The Design of Everyday Things.* Basic Books, New York, NY.

Office of Force Transformation. 2003. Operational Sense and Response Logistics: Co-Evolution of an Adaptive Enterprise Capability. Retrieved on November 8, 2007 from http://www.oft.osd.mil/library/library_files/document_229_SRL%2013-pager_%20v81.doc.

Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics. 2004. Defense Science Board study on unmanned aerial vehicles and uninhabited combat aerial vehicles. Washington, DC. Retrieved December 6, 2007 from http://www.acq.osd.mil/dsb/reports/uav.pdf.

Parasuraman, R. and Riley, V. 1997. Human and automation. Use, misuse, disuse, abuse. *Human Factors*, 39, 230–253.

Price, H., Fiorello, M., Lowry, J., Smith, M., and Kidd, J. 1980. The contribution of human factors in military system development: methodological considerations. US Army Research Institute for the Behavioral and Social Sciences. Alexandria VA. Technical Report 476.

Ripley, T. 1997. Military Communications. Retrieved November 12, 2007 from http://www.global-defence.com/1997/MilitaryCommunications.html.

Rogers, B., Palmer, B., Chitwood, J., et. al. 2004. Human-systems issues in UAV design and operation. Human Systems Information Analysis Center Wright Patterson Air Force Base, Dayton, OH. Report No.: HSIAC-RA-2004-001.

Roscoe, S. 1968. Airborne displays for flight and navigation. *Human Factors*, 10, 321–332.

Ruff, H., Narayanan, S., and Draper, M. 2002. Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned aerial vehicles. *Presence*, 11, 335–351.

Sanders, M. and McCormick, E. 1993. *Human Factors in Engineering and Design.* McGraw-Hill, New York, NY.

Sarter, N., Woods, D., and Billings, C. 1997. Automation surprises. In G. Slavendy (Ed.), *Handbook of Human Factors and Ergonomics* (2nd ed., pp. 1926–1943). Wiley, New York, NY.

Seagle, J. 1997. Unmanned aerial vehicle mishaps: a human factors analysis [Thesis]. Embry Riddle Aeronautical University Extended Campus, Norfolk, VA.

SRA International. 2006. UAV Forum. Retrieved November 8, 2007 from http://www.UAVForum.com.

Tvaryanas, A. 2006. Human systems integration in remotely piloted aircraft operations. *Aviation Space and Environmental Medicine*, 77, 1278–1282.

Tvaryanas, A., Thompson, W., and Constable, S. 2006. Human factors in remotely piloted aircraft operations: HFACS analysis of 221 mishaps over 10 years. *Aviation Space and Environmental Medicine*, 77(7), 724–732.

Turner, D. and Carstens, C. 2007. Future Force Warrior: Insights from Air Assault Expeditionary Force Assessment. Army Research Laboratory, Aberdeen Proving Ground, MD. Technical Report No.: ARL-TR-4191.

Turner, D., Carstens, C., and Torre, J. 2005. Future Force Warrior, Engineering Design Event Number 4. Army Research Laboratory, Aberdeen Proving Ground, MD. Technical Report No.: ARL-TR-3626.

US Army Natick Soldier Research, Development, and Engineering Center. 2007. Technology, Systems and Program Integration. Retrieved on November 7, 2007 from http://www.natick.army.mil/soldier/about/techprog/index.htm.

Wickens, C., Lee, J., Liu, Y., and Becker, S. 2004. *An Introduction to Human Factors Engineering*. Pearson Prentice Hall Publishing, Upper Saddle River, New Jersey.

Williams, K. n.d. A summary of unmanned aircraft accident/incident data: human factors implications. FAA Civil Aerospace Medical Institute, Oklahoma City, OK. Retrieved December 6, 2007, from http://www.hf.faa.gov/docs/508/docs/uavFY04Mishaprpt.pdf.

Wilson, J. 2002. UAVs and the human factor. *AIAA Aerospace America*, 40, 54–57.

26

Digital Warfighter Modeling for Military Applications

Karim Abdel-Malek
The University of Iowa

Jingzhou (James) Yang
Texas Tech University

Timothy Marler
The University of Iowa

Jasbir Arora
The University of Iowa

26.1 Introduction

Although human modeling capabilities and comprehensive predictive virtual humans are growing with respect to research, development, and practical application, awareness of how these tools can help design products and study humans is relatively low. If advanced enough, digital human models (DHM) provide a design aid for engineers and ergonomists. With respect to product design, such aids can help design products that are safer and easier to use. In addition, these tools can reduce costly and time-consuming production of prototypes by testing digital designs on the computer. With respect to human modeling, digital humans provide a tool with which a user can study how and why people behave the way they do. Finally, digital humans allow one to simulate real-world scenarios, with cause-and-effect studies, thus helping to save lives.

Although digital humans are dual use (military and civilian use), with potentially endless applications, one of the most significant applications is with military equipment (vehicles, weaponry, etc.). In fact, digital humans have broad applicability across all branches of the military, both with respect to human–machine interaction and with respect to individual independent human performance.

Santos™ is a premiere human modeling tool and provides an excellent example of how such tools can help the military. Thus, in this chapter, we use Santos as a platform for illustrating the many utilities and capabilities that can be incorporated in a comprehensive digital human. In addition, we demonstrate the types of problems Santos can solve with applications from a variety of military branches.

To be sure, there are many currently available human modeling tools. However, one of the most significant benefits of Santos is that he aggregates many different human modeling capabilities. He is a comprehensive DHM tool. This provides two benefits: (1) practical advantages in addressing complex human modeling problems; and (2) an ideal example of the types of capabilities that can be incorporated in a single DHM effort.

26.2 State of the Art

There is a variety of currently available human modeling tools, as well as a wide expanse of ongoing research. Because the focus of this chapter is practical applications, a brief review of the current state of the art is provided with regard to currently available tools, not ongoing research. The current human modeling tools/efforts are grouped as follows, with the company or lab indicated in parenthesis when applicable:

(1) Musculoskeletal models
- Virtual Muscle (Alfred E. Mann Institute)
- Anybody (Anybody Technology)
- SIMM (MGI Inc. and Stanford)
- NMBL (Stanford)
- LifeMod/ADAMS (Biomechanics Research Group, Inc.)

(2) Rigid-body dynamics models
- LifeMod/ADAMS (Biomechanics Research Group, Inc.)
- Articulated Total Body Model (AFRL)
- Madymo (tass)

(3) DHM with motion
- Dhaiba (DHRC in Japan)
- DI-Guy and Digital Biomechanics (Boston Dynamics)
- Jack (Siemens) and Crew Station (Micro Analysis and Design)
- Ramsis (Human Solutions)
- Safework Pro (Safework Inc./Dassault Systems)
- Santos (The University of Iowa, VSR)

(4) DHM without motion
- Boeing Human Modeling System (BHMS)

(5) Other

Robot:
- Asimo (Honda)
- *Motion-capture system*:
- Motek

Ergonomics method:
- Owas (Tampere University of Technology)

Internal organs:
- Virtual Soldier Project (DARPA)

Regarding categories 3 and 4, although DHM is a general term with broad applicability, in this context, it refers to complete human models that actually look like real humans to some extent. BHMS is a DHM that acts purely as a mannequin that must be manipulated by the user. Other DHMs can provide motion prediction and/or simulation to some extent. Most DHMs that actually predict motion do so based on prerecorded motion-capture data. Digital biomechanics then incorporates optimal controls to determine how the model gravitates to the prerecorded motion, given additional input. Of the DHMs with motion, Safework Pro has the most limited capabilities, as motion is simply based on interpolation between given postures.

Some general conclusions are drawn based on a comparison of the above-mentioned tools. In general, DHMs do not represent the current state of the art when compared to ongoing research. That is, the currently available products are not as advanced as they could be. Musculoskeletal models, however, are relatively advanced. Most DHMs are geared toward cab/cockpit evaluation and posture analysis. Less functionality is available that targets actually studying the human in general scenarios. Basic skeletal structures are lacking; they are too simple. There is relatively little work that incorporates accurate hand models and fatigue (muscle-based or cardiovascular).

Santos is responding to these deficiencies in the current state of the art. In particular, relative to currently available tools, the following distinct strengths stand out with Santos:

- It is a comprehensive model that incorporates a wide variety of capabilities.
- Its fidelity, with respect to appearance and skeletal structure, is superior.
- It incorporates the most extensive hand model.
- It is truly predictive, with no need for prerecorded data or movies, when simulating human posture and motion.

Regarding predictive capabilities, *technically*, all of the DHMs with motion are predictive to some extent, and *technically*, they are all data driven in the sense that any model requires input data. However, Santos is the only DHM that does not require actual motion data in order to predict motion. The input for Santos is human model information, not posture or motion information. Thus, it is the only model that truly predicts posture and motion.

26.3 Santos Capabilities

Although many efforts to model independent facets of the human body are mature, especially in the medical field, there has been relatively little advancement in modeling complete humans and how they interact with products. The tools that are currently available for the latter thrust are limited primarily in terms of their fidelity and autonomy. Santos is a new kind of virtual human (Figure 26.1), developed at VSR, and he responds to these limitations.

Santos's current strengths are not only in individual capabilities but also in the combination of and connectivity of capabilities. In short, Santos is unique in its combination of superior features. The core functionality for any virtual human is the ability to model human posture and motion realistically. In this respect, Santos is *predictive*. Posture and motion are not based on prerecorded motion or data that must be updated when there are alterations in the virtual environment. Rather, Santos actually predicts how people strike various poses and how they move. This allows one to study, for example, changes in gait when there are changes made to the model such as backpack weight, armor characteristics, physiological indices, or strength characteristics. Such predictive capabilities afford the model considerable autonomy and stem, in large part, from our *optimization-based* approach to human modeling. The joint angles (one for each degree of freedom) in the Santos model provide the design variables that are determined by minimizing human performance measures such as speed, energy, and fatigue. This computational optimization problem is solved while satisfying constraints that are used to model various tasks. For instance, limits on joint angles can be imposed based on anthropometric data. Our optimization-based approach allows us considerable predictive autonomy. As different environmental conditions are imposed, no new data is needed. Our approach is also extremely fast, allowing Santos to operate and react in *real time* in most cases. The most unique aspect of our approach is the ability to handle complex dynamics problems. Equations of motion are simply incorporated as a constraint in the optimization problem, thus allowing us to conduct inverse and forward dynamics simultaneously, without having to manage potentially slow and intractable time-integration algorithms. We call this *predictive dynamics*. Along with joint angles that dictate predicted motion, we are able to use dynamic analysis to recover actuation torques (at each joint) necessary to cause motion. The performance measures, which serve as objective functions in the optimization problem, govern the motion and posture. However, what dictates one's motion varies depending on the task being completed. With our approach, we are able to exchange and combine performance measures, thus modeling different kinds of behavior with what we

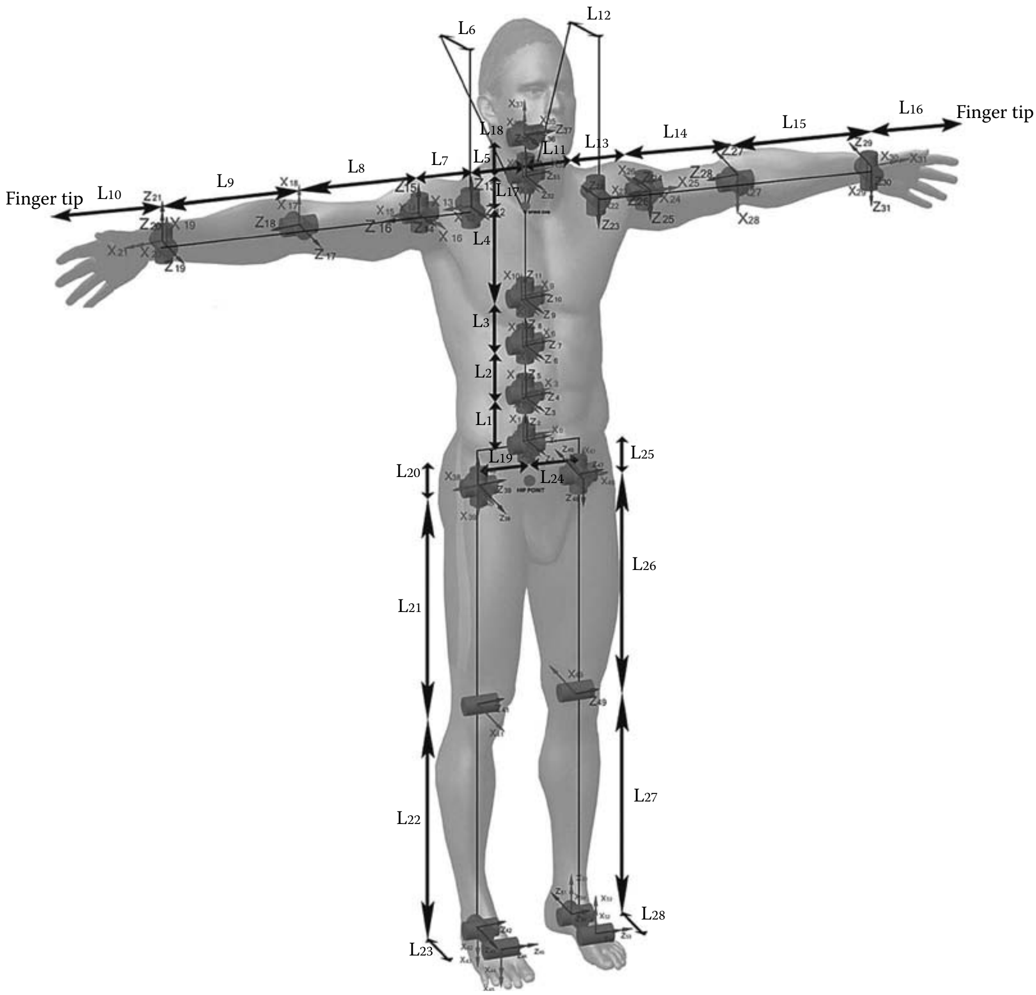


FIGURE 26.1 Santos' kinematic model.

call *task-based* prediction. Because of our use of optimization, the number of degrees of freedom (DOF) used for the human model does not significantly impact computational speed. Thus, our skeletal model currently includes 209 degrees of freedom, and it can easily be expanded without significant computational detriment. This model for Santos is extremely flexible, in that it can be modified to represent any anthropometric cross section. The link lengths that represent dimensions of different bones can be modified on the fly, as can joint limits. The location of different joints can also be altered. The *high fidelity* of the Santos model is also reflected in its appearance. Santos is not a cartoon. His realism is superior, and this allows Santos to act as an actual design aid, a virtual human that can ultimately train and work with real humans. In this section, we summarize some of Santos' capabilities, which are listed in Figure 26.2.

26.3.1 Posture and Motion

A key capability for Santos is a novel optimization-based approach to posture and motion prediction (Marler 2005; Yang et al. 2006; Xiang et al. 2007; Abdel-Malek et al. 2008). With this approach, joint angles provide design variables, constrained by joint limits based on anthropometric data as well as

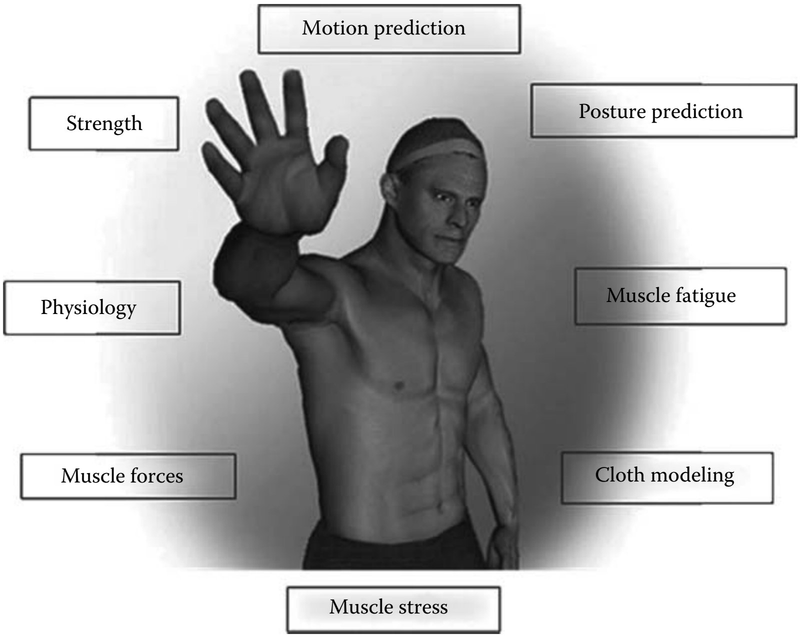


FIGURE 26.2 Capabilities.

other physical constraints. Human performance measures that represent physically significant quantities, such as energy, discomfort, etc. provide the objective functions. Using this approach affords the digital human a substantial amount of autonomy, thus enabling an avatar to react to infinitely many scenarios rather than having to draw on prerecorded motions. By finding the joint angles that optimize a performance measure, one can determine the most appropriate posture for the human body, which is highly redundant with respect to feasible postures and motions. With the optimization approach, incorporating additional capabilities or features is usually a matter of simply introducing new constraints and/or objective functions. In addition, this optimization-based approach allows us to operate in real time (Figure 26.3).

Within posture prediction capability, the *zone differentiation* tool (Yang et al. 2008; Yang et al. in press) allows the user to visualize not only the complete reach envelope but also the interior levels of the envelope in terms of various performance measures. It uses a color map to display the relative values of the performance measures (i.e. comfort, joint displacement) at points surrounding an avatar. Using this tool, a vehicle designer can visually display the impact that the placement of a control (switch, button, etc.) has on a driver's postural comfort. The comfort values are displayed in a manner similar to how finite element analysis (FEA) programs display stress and strain results. The development of this tool requires two main components. First, both the simulated postures and the resultant comfort levels are correlated against actual experimental results. Second, the software tools needed to calculate and display the comfort zones, as well as the graphical user interface, are developed.

As an extension of posture prediction, optimization-based motion prediction is called predictive dynamics (Kim et al. 2005; Kim et al. 2006). It is an approach for predicting and simulating human motion. This approach avoids solving typical differential algebraic equations (or ordinary differential equations) in order to create the resulting simulations for highly redundant systems. Detailed and anatomically correct joint-based full-body human models with high degrees of freedom can thus be used to create more realistic simulation of tasks with relatively less computation. Various tasks like walking (Xiang et al. 2007), running (Chung et al. 2007), stair climbing, throwing, and box lifting (Abdel-Malek et al. 2008) have been simulated using this approach.

FIGURE 26.3 Aiming.

26.3.2 Strength

In vivo muscle force is a highly nonlinear phenomenon that is dependent on factors such as muscle length, contraction velocity, and past contractile history (e.g. fatigue). While muscle force has been modeled as a linear system, linear representations are not as accurate as more complex nonlinear models (Frey Law and Shields 2006). Force decays nonlinearly with increasing velocity (Hill 1938). Active muscle force varies with muscle length, due to varying overlap of the force-producing filaments (i.e. actin and myosin) and stretching of structural proteins at long muscle lengths. It is not a simple transformation to apply these principles from the single muscle level to the joint level due to multiple synergistic muscles acting at a joint and the varying muscle moment arms with joint angle. Our approach to modeling joint strength inherently includes each of these nonlinearities. We are experimentally measuring joint peak torque at several angles through the normal range of motion and at several angular velocities (e.g. 0–300°/sec) for six major body joints. These data are used to create 3D surfaces, with peak torque (strength) as a function of joint position and angular velocity (Laake and Frey Law 2007) While previous authors have reported similar 3D representations of joint strength (Anderson et al. 2007), no one has developed a normative database to use for DHM. This database will allow us to represent human strength capability as percentiles, similar to how anthropometry is represented, e.g. 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of human strength capability for men and women. Note that these population percentiles are unrelated to anthropometric (height) percentiles, as tall individuals can be relatively weak or, conversely, short individuals relatively strong. Santos predicts the required joint torques (versus time), along with joint angle and velocity needed to accomplish a given dynamic task. If we plot the predicted joint torque versus joint angle and angular velocity at each point in time, we can assess the magnitude of the predicted joint torque relative to a population percentile's maximum capability (i.e. 50th percentile male normative 3D surface). This provides a model of percent effort, where the closer the predicted task lies to the maximum surface, the more difficult a task becomes. This provides a unique methodology for a digital human to predict the perceived level of difficulty of a subtask, incorporating known muscle force determinants such as muscle length, moment arm, and shortening velocity into one simple process in Figure 26.4 (Abdel-Malek et al. 2008).

26.3.3 Fatigue

As muscles fatigue, both maximum torque and velocity are impacted, e.g. it becomes increasingly more difficult to generate large joint torques and high movement velocities. Thus, we can use the 3D joint strength surfaces to represent fatigue by decaying them with repetitive or higher intensity activities. We have developed a model that predicts how a joint surface will decay over time, using a series of differential equations based on compartment and control theories (Xia and Frey Law 2008). A single three-compartment model represents muscles involved at a joint in one of three states: active, resting, or fatigued. Rate constants define the behavior of the transfer between the active and fatigued compartments; however, we use a proportional controller to define the transfer between resting and active states. The model determines how much of the resting muscle pool must be activated in order to match the predicted joint torques. The combined size of the resting and active pools determines the residual capacity of the system for use as a decay coefficient (values between 0 and 1) to decay the 3D strength surface. This can be used both as a feedback, post-processing mechanism, providing a means to measure "time to fatigue" when the task is no longer feasible without alterations in the predicted dynamics, and/or as a feed-forward mechanism where the predicted dynamics can change as the available strength decreases. Most notably, we incorporate strength nonlinearities into our fatigue model by normalizing predicted dynamic joint torques by the corresponding peak 3D strength representation (percent of maximum

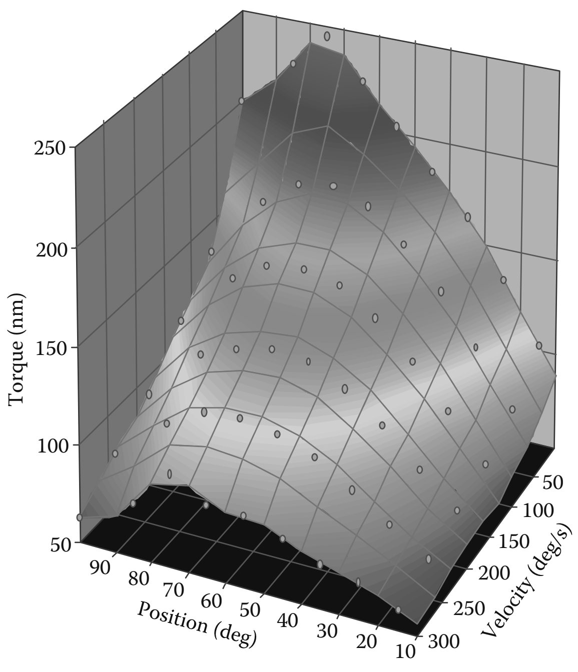


FIGURE 26.4 Mean male knee extension, for 50 percentile strength.

torque). Thus, the model targets this percent max torque rather than a specific absolute muscle force or torque, as typically used by other models.

26.3.4 Muscle Forces and Stress

In addition to modeling muscle restrictions in terms of joints, we are also conducting research on how to predict muscle activation using optimization in Figure 26.5 (Patrick 2005). Different muscle group combinations can generate the same motion. Muscle injuries are common in many different motions and because the muscle force exceeds its limit. An optimization-based approach has been developed for Santos to determine the muscle forces. Minimizing muscles' activities is the human performance measure used as the criterion to solve this redundant actuation system subject to muscle forces within their limits. During human movement, muscles are deformed and muscle stress is a key factor to determine

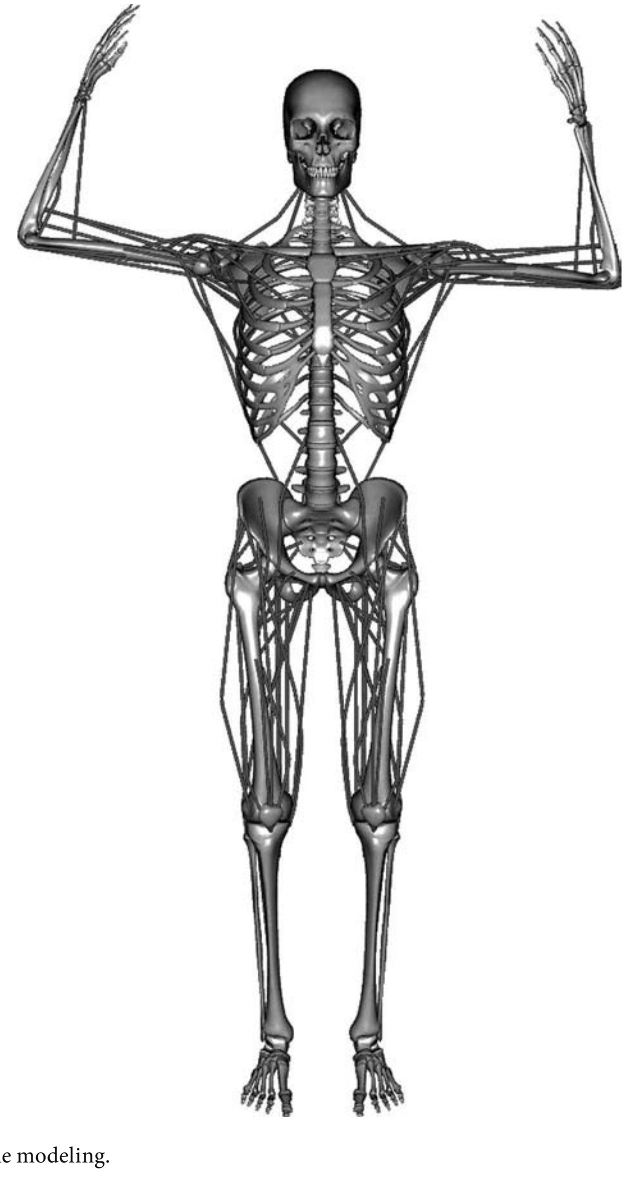


FIGURE 26.5 Muscle modeling.

the biomechanic characteristics of muscles. A NURBS-FEM real-time method for calculating muscle deformations and stress has been developed (Zhou and Lu 2005). Salient features of the method include: (1) it provides a better (smoother) geometric description; and (2) it involves fewer degrees of freedom. The method by itself is sufficiently general, applicable to general deformable bodies. Muscle motion can be either active or passive. Active muscle generates forces inside the muscle by fiber contraction. The degree of fiber contraction is controlled by the neural input that is mathematically represented by a scalar parameter in our constitutive model. If this scalar parameter equals zero, the muscle is fully passive and does not generate any contractive force. There are two types of major contraction of muscle. One is isometric contraction, where the muscle contracts or tenses without changing its length; the other is isotonic contraction, where the length of a muscle changes while keeping the contractive force constant.

26.3.5 Cloth Modeling

Santos' clothing is modeled as flexible continuum shells draped onto an active piecewise rigid human body surface (Figure 26.6). The clothing model undergoes unilateral frictional contact with the moving human body model and exerts associated forces on the human body. These forces can then be used

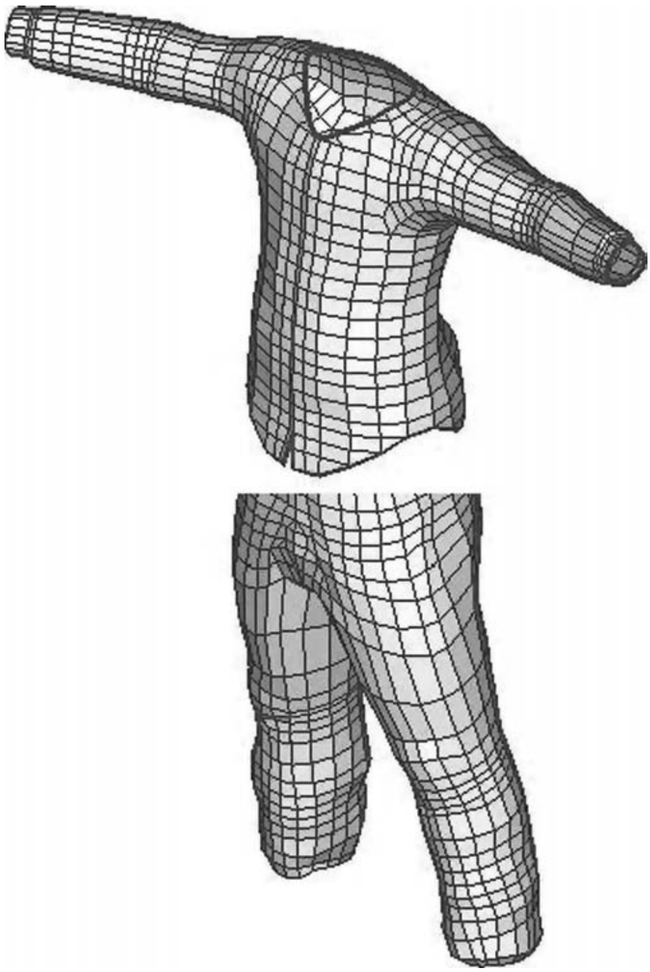


FIGURE 26.6 Clothing.

to calculate the energy necessary to move the clothing, which can be substantial with heavy protective suites. Here, we describe how a subject-specific human body surface driven by predictive dynamics is approximated with piecewise rigid mesh segments and how the clothing is draped onto the body surface. Although not the focus here, constitutive behavior of different clothing fabrics is an important aspect of clothing modeling (Swan et al. 2007). The starting point for a virtual mannequin representing the anthropometry of a specific human subject is a laser body scan, which yields a polygonal mesh (Ashdown 2007) involving hundreds of thousands of nodes and polygons. Such meshes are much finer than is necessary for clothing modeling, and indeed usage of such fine meshes would be too computationally expensive in clothing contact modeling. Accordingly, the body-surface mesh can be coarsened using commercial software tools. Once a complete and coarsened body scan mesh is obtained, it is decomposed into an assemblage of individual meshes corresponding to limb or torso segments. The optimal decomposition of meshes into segments can itself be quite involved (see, e.g. Lien 2006) but was done in an ad-hoc manner at the major joint locations using AutoCAD. The individual body mesh segments are subjected to rigid body translations and rotations in time to approximate evolution of the body surface as the human subject performs physical tasks. Geometric inconsistencies will develop at the joints between the rigid body mesh segments, and since such gaps present a problem in clothing modeling, they are patched in the current framework using auxiliary spherical and ellipsoidal rigid mesh segments at each of the joints between body mesh segments. With a controllable piecewise rigid body surface for the mannequin in place, clothing models can be draped onto the mannequin. A number of different approaches have been taken to get the clothing model onto the human body. Two in particular are trying to simulate the actual dressing process (Man and Swan, 2007; Volino and Magnenat-Thalmann, 2000) and trying to preposition models of clothing patterns around models of the human body and then bring the patterns together and stitch them up at their seams (Grob et al. 2003), essentially constructing the garment about the mannequin. Here, we try a different approach that involves two steps: (1) the clothing is statically prepositioned over the mannequin without any concern for contact or penetration between the body and the clothing; and (2) once the clothing model is in place, contact mechanics (penetration, detection, and correction) are turned on to eliminate clothing penetrations of the body surface. When prepositioning the assembled clothing models about the body in a fixed posture, the objective is to bring the centroids of the clothing model edges (cuffs, waistline, necklines, etc.) into alignment with the corresponding body-segment edge centroids. This is achieved by using penalty forces that are significant when the clothing and body segment edge centroids are not coincident and are minimized as the clothing becomes properly positioned on the body. This prepositioning problem is solved quasi-statically, leaving the clothing model properly positioned on the body in a gross sense, although with some significant clothing penetrations of the body. In the second stage, explicit dynamic analysis of the clothing model is performed while gravity loading pulls downward on the clothing model, and explicit contact analysis as described by Man and Swan (2008) is utilized to remove starting penetrations and any other penetrations that develop. Virtually all of the initial clothing penetrations from the prepositioning stage are eliminated in this phase of analysis. From this stage, the mannequin can be activated based on joint-angle profiles from predictive dynamics, and the clothing model will respond accordingly via frictional contact interactions with the mannequin surface.

26.3.6 Physiology

A process has been developed at VSR whereby existing models for various physiological quantities have been linked and incorporated in a digital human (Mathai 2005; Yang et al. 2005). The main physiological indices include oxygen uptake and heart rate. As work increases (Figure 26.7), blood circulation increases to deliver more oxygen to the muscles, which in turn increases the ventilation rate to supply oxygen to the blood. The detailed oxygen uptake formulation is provided by Mathai (2005). Given the oxygen uptake, it is possible to model the heart rate. The relationship between heart rate (HR)

FIGURE 26.7 Physiological output.

and VO_2 is linear (Wilmore and Haskell 1971). As activity levels increase, the heart rate increases as well. The increase in heart rate is in response to the increased need for oxygen to generate ATP by the muscles involved in activities. By beating faster, the heart pumps more blood to the muscles. Although there are additional functions of increased blood flow, they are relatively minor compared to the needs of the muscles and thus may be neglected. There are three mechanisms by which the body supplies increased oxygen demands to the working muscles: arterio-venous oxygen difference, increased stroke volume, and increased heart rate (Astrand and Rodahl 1970). The heart rate equation is described by Mathai (2005). The most significant repercussion of this model for heart rate and oxygen uptake is the ability to approximate overall exhaustion. The simulated oxygen uptake is compared to general guidelines for maximum oxygen uptake (determined experimentally) to give a general indication of overall exhaustion.

26.4 Military Applications

26.4.1 Cockpit Design–Air Force–Posture Prediction

Posture analysis tools are often used in cab and cockpit design, primarily for reach analysis yielding envelopes. In addition, avatars that act as digital mannequins can be positioned by an expert user to represent a realistic posture that is subsequently evaluated with respect to comfort and other ergonomic indices. However, advanced posture-prediction tools, such as those incorporated in Santos, can provide additional capabilities.

Because of the way optimization-based posture prediction works, one can output the value of various performance measures like discomfort, joint displacement, potential energy, effort, and vision indices (Figure 26.8). This allows one to evaluate different postures and thus different designs that require the predicted postures. In addition, by combining various performance measures using multiobjective optimization, one can study which factors drive human posture under various conditions and scenarios. These features yield a powerful reach-analysis tool that can be applied to cockpit design.

Furthermore, because optimization-based posture prediction is extremely fast, one can essentially use it for real-time inverse kinematics. That is, the user can drag any point on Santos to any other point in space, and all joint angles are automatically determined/predicted in real time, rather than having the user set each joint angle (Figure 26.9). This saves a substantial amount of developmental time when positioning an avatar in a virtual environment for ergonomic studies.

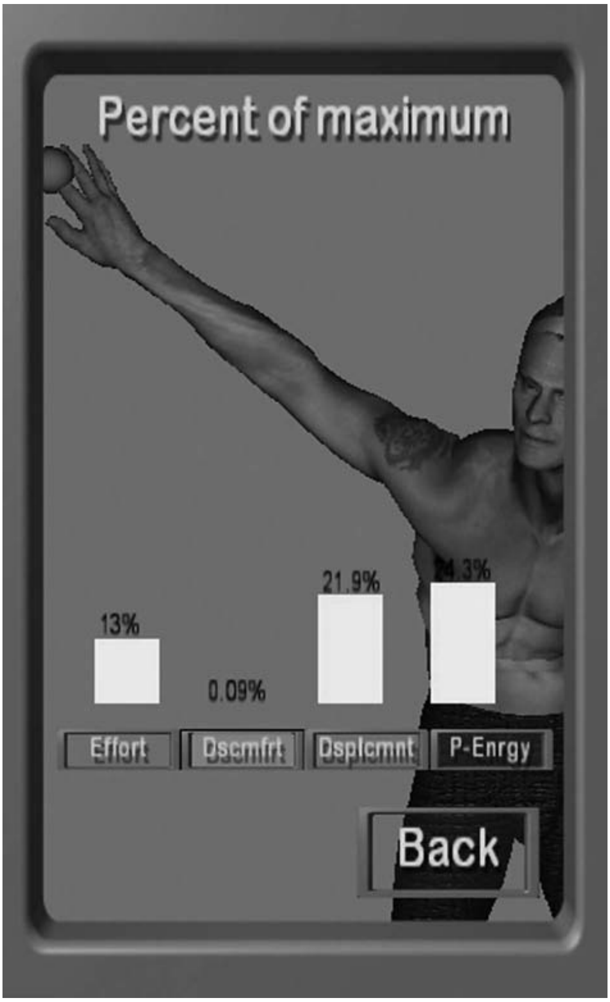


FIGURE 26.8 Human performance measures.

Finally, the zone differentiation tool discussed in Section 26.3 can be used to analyze a preexisting cab design. The 3D color contours developed using zone differentiation produce a cloud of points, with each point representing a different performance-measure value (Figure 26.10). This cloud can be manipulated within the virtual environment and used to colorize geometry in the environment, as shown in Figure 26.11.

26.4.2 Tank Ingress/Egress–Army–Motion Prediction, Muscle Force/Stress

In the last section, we discussed a digital warfighter's application in cockpit design. In this example, we demonstrate another application for the digital warfighter in tank design. There are three areas in which the digital warfighter is used to evaluate the design: (1) tank interior design; (2) design strategies for ingress/egress; (3) suspension and seat system design; and (4) strength and fatigue assessment. Regarding tank interior design, it is similar to cockpit design. Posture prediction and zone differentiation are used to check the layout design of controls and buttons within the tank. An ergonomic suit is used to evaluate suspension and seat systems. When this digital human drives the tank on different

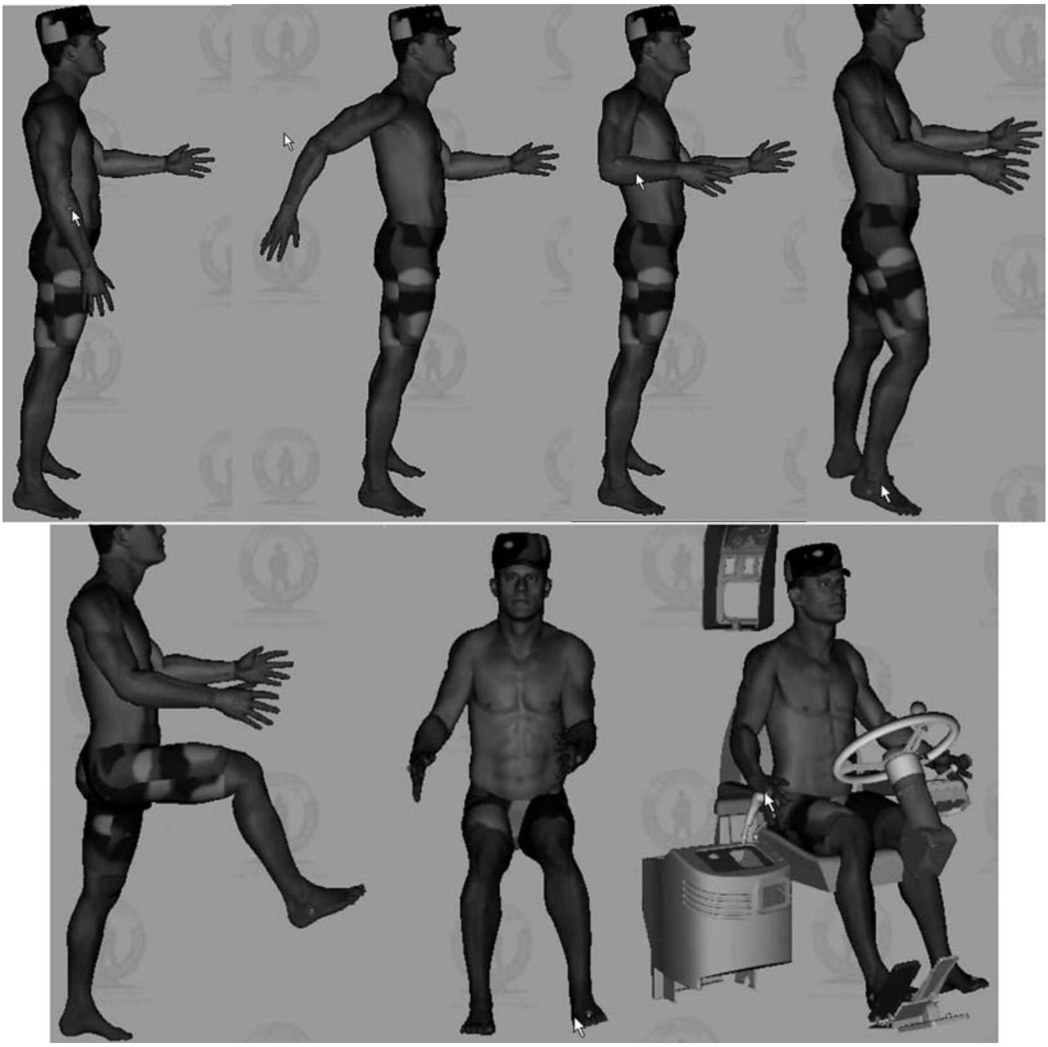


FIGURE 26.9 Real-time inverse kinematics.

terrains at a certain speed for one hour, we can assess the strength and fatigue levels for the digital human. In this section, we mainly focus on the design strategies for ingress and egress.

Ingress/egress is an important factor that has to be carefully considered in the design process of a tank. The major reason is to reduce injuries of tank drivers and improve the ergonomic design. Figure 26.12 shows snapshots of ingress/egress simulation using predictive dynamics. There are many ways (paths) for the tank driver to get in and out of the tank. We can simulate all different cases and compare the maximum joint torques in the whole body. Based on the joint torques, we can select the best path and strategy for ingress/egress during the design stage.

26.4.3 Independent Tasks–Marines–Physiology, Cloth, and Armor

In addition to studying man–machine interactions for product design, digital humans can be used to study actual human motion during tasks like running, diving, and aiming (Figure 26.13). These tasks are

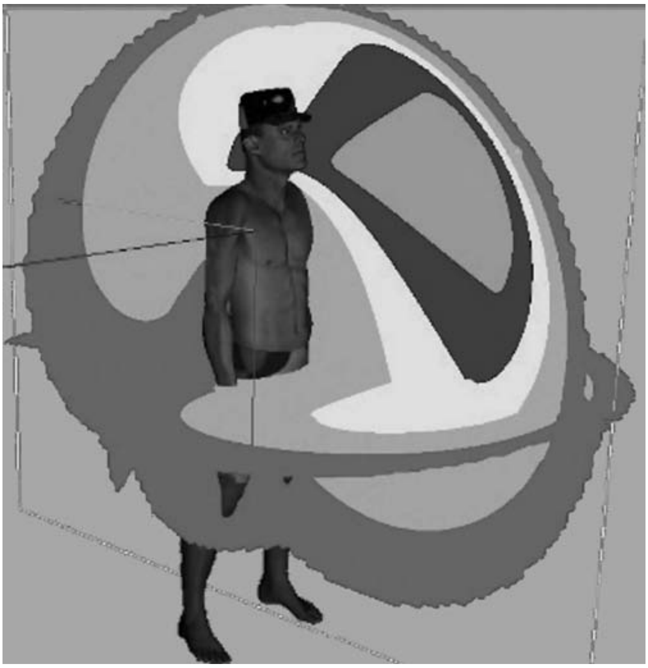


FIGURE 26.10 Cloud of points for zone differentiation.

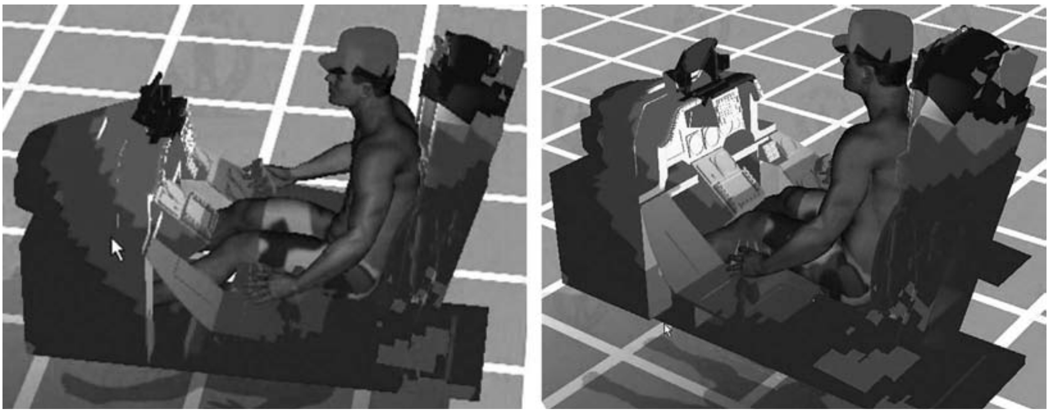


FIGURE 26.11 Using zone differentiation to colorize cockpit geometry.

simulated using predictive dynamics (dynamic motion prediction). Concurrently, the user can extract information like joint angle profiles, joint torque profiles, and ground reaction forces (Figure 26.14).

Using this feedback, the user can conduct if-then studies to see the effects of changes in anthropometry or external loads (i.e. back pack weight). It is also possible to see the effects on changes in joint ranges of motion induced by body armor, as shown in Figure 26.15. The user can design various body armor systems, place them on the avatar, and see how such armor affects tasks like running, diving, and aiming.

One of the advantages of a comprehensive DHM is the ability to link different modeling capabilities, as shown in Figure 26.16. In fact, motion prediction is central to any human modeling tool; most other capabilities either affect motion or are affected by motion. For instance, clothing models predict both the

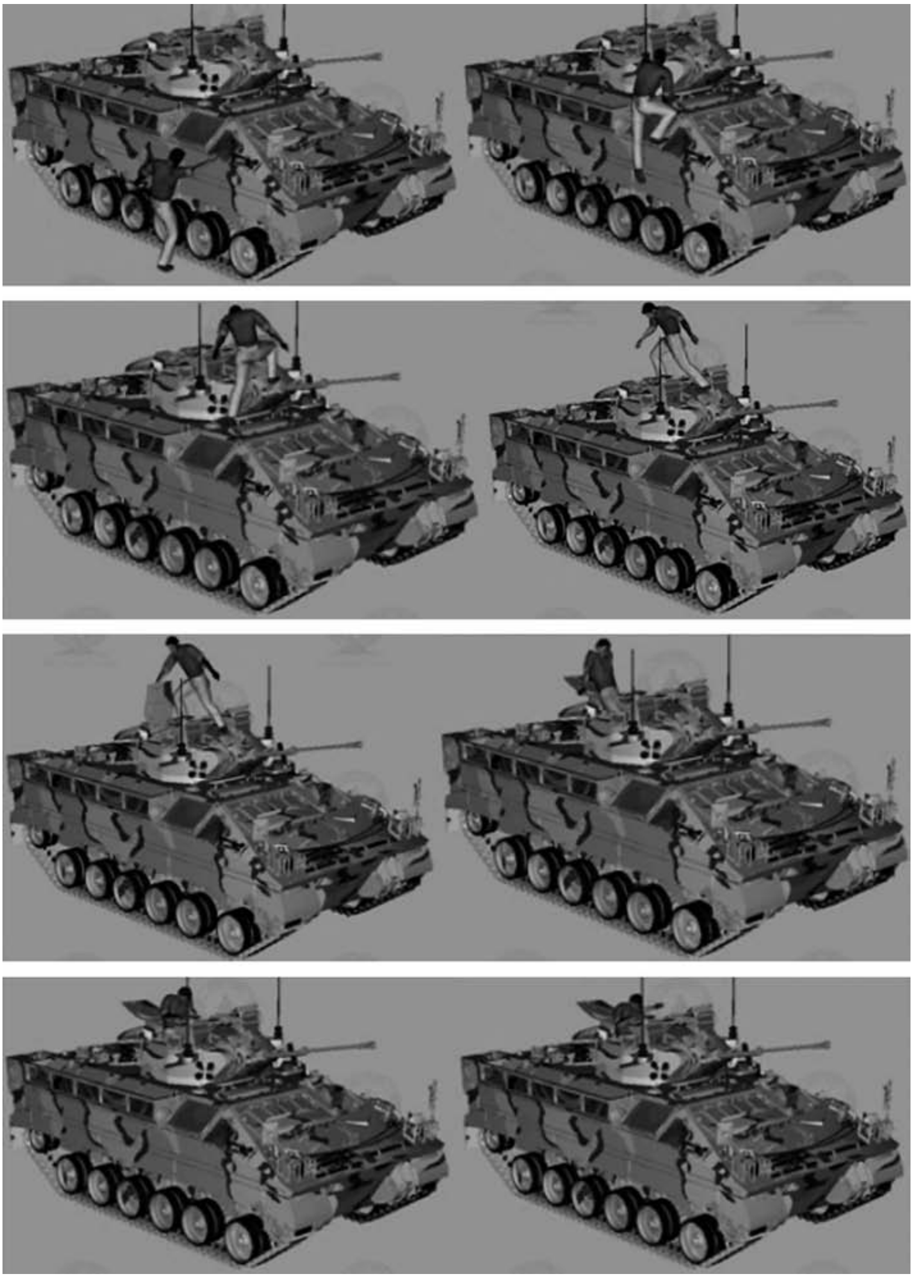


FIGURE 26.12 Snapshots of ingress/igress simulation using predictive dynamics.

energy necessary to carry heavy clothing (i.e. protective suits) as well as the additional joint torque necessary to bend such clothing. The additional joint torque can then be incorporated in predictive dynamics as constraints, and the necessary energy can be added to the performance measure. As with armor modeling, the user can conduct if-then studies using various types of clothing or protective garbs.

Furthermore, values from an energy-based performance measure can be extracted and used to calculate various physiological indices, as discussed in Section 26.3. Then, a user can study how avatars with various characteristics (anthropometry, physical conditioning, etc.) perform tasks differently.

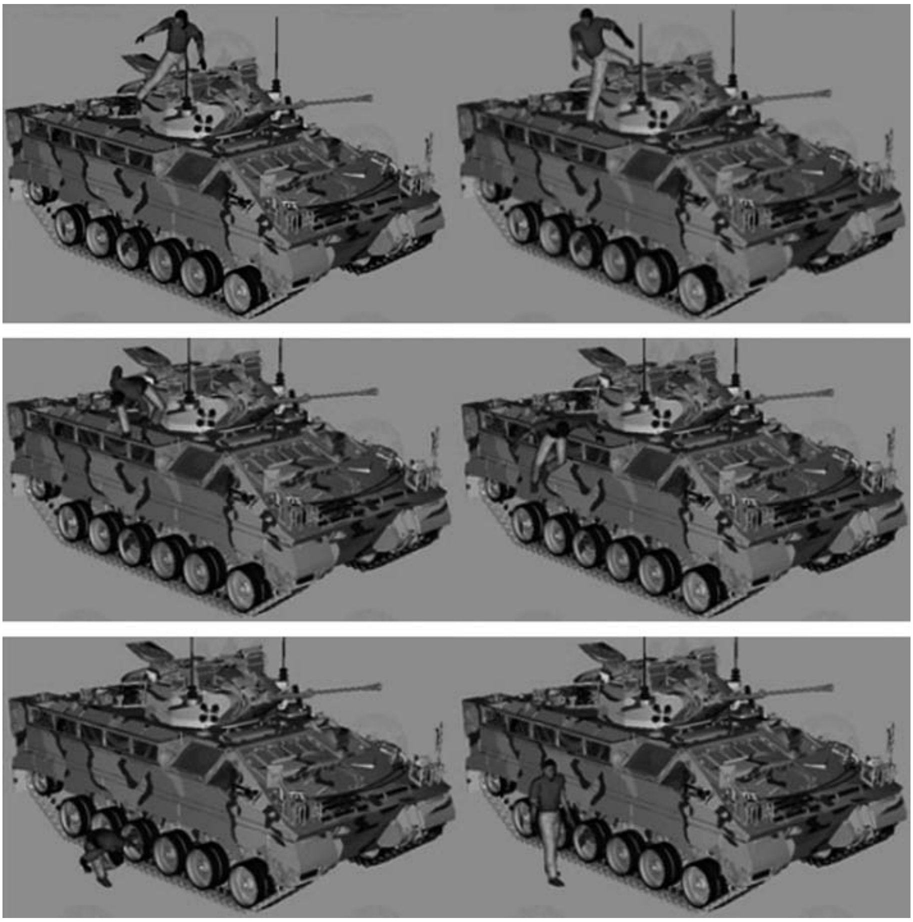


FIGURE 26.12 (Continued)

FIGURE 26.13 Running, diving, and aiming.

26.4.4 Loading Shells–Navy–Strength/Fatigue

In this application example, we illustrate how the DHM can help to design the shell-loading system and the strategy to load the shells in a Navy ship. In the Navy ship design, one key component is the system of shell-loading system. The goal is to reduce soldiers' injuries and fatigue levels during loading shells through a better design of the loading system. Figure 26.17 is a snapshot of the digital human loading the shell using predictive dynamics. We can check joint torques and fatigue levels to determine the loading frequency, time for shift, and the loading system design.

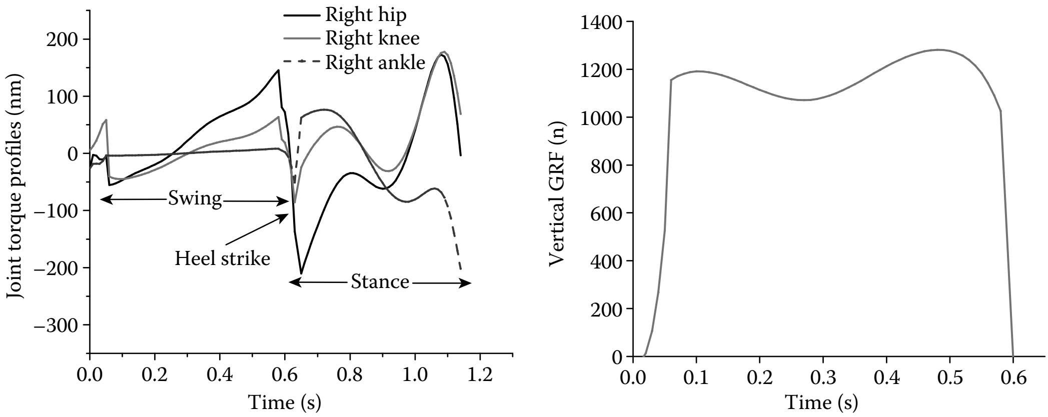


FIGURE 26.14 Joint torque profile and ground reaction forces (GRF).

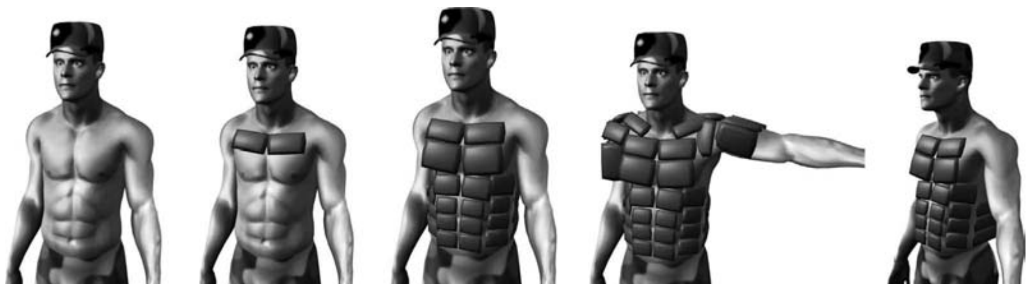


FIGURE 26.15 Body armor.

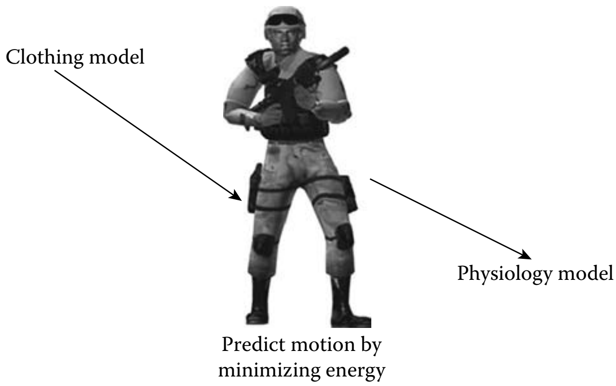


FIGURE 26.16 Linking human-modeling capabilities.

26.5 Conclusions

This chapter has used Santos, a comprehensive and advanced DHM, to present the kinds of capabilities and tools that human models can provide. In addition, it has shown how a DHM can be used with a variety of military applications to solve different kinds of problems.

Especially when considering military applications, almost all tools, vehicles, and equipment require human interaction. This interaction must be considered during the design process, in order to maximize

FIGURE 26.17 Snapshot of the digital human loading the shell using predictive dynamics.

efficiency and safety while minimizing cost. DHMs provide an effective means for incorporating the human element during product design. In addition, DHMs provide a unique tool for studying how and why humans behave the way they do.

Clearly, the human system is complex and multidisciplinary. Consequently, providing a high-fidelity avatar requires multiscale efforts drawing from a variety of fields and disciplines. Often, researchers do not realize that their work is applicable to human modeling, but in this chapter, we have demonstrated how areas as diverse as biomechanics, computer science, physiology, optimization, robotics, and others can contribute to the field of human modeling.

References

Abdel-Malek, K., Arora, J., Frey Law, L., Swan, C., Beck, S., Xia, T., Bhatt, R., Kim, J., Xiang, Y., Rasmussen, M., Murphy, C., Laake, A., Marler, T., Yang, J., and Obusek, J. 2008. Santos: A digital human in the making. The IASTED International Conference on Applied Simulation and Modelling (ASM08), Corfu, Greece, June 23–25.

Anderson, D.E., Madigan, M.L., Nussbaum, M.A. 2007. "Maximum voluntary joint torque as a function of joint angle and angular velocity: Model development and application to the lower limb." *Journal of Biomechanics*, 40(14), 3105–3113.

Ashdown, S. 2007. Available online at http://www.bodyscan.human.cornell.edu.

Åstrand, P.-O., and Rodahl, K. 1970. *Textbook of Work Physiology*. McGraw Hill, New York.

Chung, H.J., Xiang, Y., Mathai, A., Rahmatalla, S., Kim, J., Marler, T., Beck, S., Yang, J., Arora, J., and Abdel-Malek, K. 2007. A robust numerical formulation for the prediction of human running. Proceedings of SAE Digital Human Modeling for Design and Engineering, Seattle, June 12–14. University of Washington, WA. Society of Automotive Engineers, Warrendale, PA.

Frey Law, L.A., and Shields, R.K. 2006. Predicting human chronically paralyzed muscle force: A comparison of three mathematical models. *Journal of Applied Physiology*, 100, 1027–1036.

Grob, C., Fuhrmann, A., and Luckas, V. 2003. Pre-positioning of virtual clothing. *Proceedings 19th Spring Conference on Computer Graphics*. ACM. New York, NY.

Hill, V.A. 1938. The heat of shortening and the dynamic constants of muscle. *Proceedings of the Royal Society of London B*, 126, 136–195.

Kim, H.J., Horn, E., Arora, J., and Abdel-Malek, K. 2005. An optimization-based methodology to predict digital human gait motion. Proceedings of SAE Digital Human Modeling for Design and Engineering, Iowa City, IA, June 14–16.

Kim, J., Abdel-Malek, K., Yang, J., and Marler, T. 2006. Prediction and analysis of human motion dynamics performing various tasks. *International Journal of Human Factors Modelling and Simulation*, 1(1), 69–94.

Laake, A., and Frey Law, L.A. 2007. Modeling 3D knee torque surfaces for males and females. Presented at American Society for Biomechanics, Palo Alto, CA.

Lien, J-M. 2006. *Approximate convex decomposition and its applications*, PhD Dissertation, Texas A&M University.

Man, X., and Swan, C.C. 2007. A mathematical modeling framework for analysis of functional clothing. *Journal of Engineered Fibers and Fabrics*, 2(3), 10–28.

Mathai, A. 2005. Towards physiological modeling in virtual humans. M.S. Thesis, The University of Iowa, Iowa City, IA.

Marler, R.T. 2005. A study of multi-objective optimization methods for engineering applications. Ph.D. Dissertation, University of Iowa, Iowa City, IA.

Patrick, A. 2005. "Development of a 3D model of the human arm for real-time interaction and muscle activation prediction." MS Thesis, The University of Iowa, December 2005.

Swan, C.C., Man, X., Yang, J., Rasmussen, M., and Williams, R. 2007. Multiscale analysis for virtual prototyping and performance testing of protective fabrics. Paper IMECE2007-43925 presented at ASME IMECE, Seattle, WA.

Volino, P., and Magnenat-Thalmann, N. 2000. *Virtual Clothing, Theory and Practice*. Springer, Berlin.

Wilmore, J.H., and Haskell, W.L. 1971. Use of the heart rate-energy expenditure relationship in the individualized prescription of exercise. *American Journal of Clinical Nutrition*, 24, 1186–1192.

Xia T., and Frey Law L. 2008. Multiscale approach to muscle fatigue modeling. Pacific Symposium on Biocomputation, January 4–8, Big Island, HI.

Xiang, Y., Chung, H.J., Mathai, A., Rahmatalla, S., Kim, J., Marler, T., Beck, S., Yang, J., Arora, J., Abdel-Malek, K., and Obusek, J. 2007. "Optimization-based Dynamic Human Walking Prediction." *Proceedings of SAE Digital Human Modeling for Design and Engineering*, June 12–14, Seattle, University of Washington, WA, USA.

Yang, J., Marler, R.T., Beck, S., Abdel-Malek, K, and Kim, J. 2006. Real-time optimal reach-posture prediction in a new interactive virtual environment. *Journal of Computer Science and Technology*, 21(2), 189–198.

Yang, J., Marler, R.T., Kim, H.J., Farrell, K., Mathai, A., Beck, S., Abdel-Malek, K., Arora, J., and Nebel, K. 2005. Santos: A new generation of virtual humans. *SAE 2005 World Congress, April, Detroit, MI*. Society of Automotive Engineers, Warrendale, PA.

Yang, J., Sinokrot, T., and Abdel-Malek, K. 2008. "Workspace Zone Differentiation Analysis and Visualization for Virtual Humans." *Ergonomics*, 51(3), 395–413.

Yang, J., Verma, U., Marler, T., Beck, S., Rahmatalla, S., and Harrison, C. "Workspace Zone Differentiation Tool for Visualization of Seated Postural Comfort." *International Journal of Industrial Ergonomics*. (In press)

Zhou, X., and Lu, J. 2005. NURBS-based Galerkin method and application to skeletal muscle modeling. *Proceedings of the 2005 ACM Symposium on Solid and Physical Modeling Cambridge, SPM '05*. ACM Press, New York, NY, 71–78.

VII

Management and Process Improvement

27

Achieving Strategic Aims: Moving Toward a Process-Based Government Enterprise

George R. Freeman
Air Force Institute of Technology

Since the end of the Cold War, the Unites States Department of Defense (DoD) has been increasingly called upon to perform everything from toppling standing governments, to nation building and everything in-between. For almost two decades the (DoD) has struggled to meet these divergent demands with limited resources. Increasingly, the DoD recognizes that a new way of developing, fielding, employing and sustaining capabilities is needed as current requirements are creating resource demands and tasking the DoD beyond its breaking point.

The question becomes, how does the DoD chart a course, which when pursued, will yield an organizational construct and associated strategic alignment and execution process which simultaneously meets these very divergent needs at the lowest possible cost to the taxpayer? And more importantly, can it rapidly re-direct finite resources to achieve the desired strategic effect?

This chapter will present various theories, and introduce the reader to a model which, when properly employed, aligns an enterprise to most optimally meet customer requirements. You will learn a new way for DoD to look at itself—from their customer's perspective, and from that view better conceptualize a new organizational construct and associated strategic planning and execution process which takes the DoD beyond historical stovepipes and simultaneously successfully links strategic aims to measurable tasks.

In looking at the DoD, especially given the events of the last several years, one is left asking; what business is the DoD really in? As demands rapidly change is the DoD agile enough to rapidly respond to changing and often fluid strategic aims? How do members of the DoD rapidly adjust their resources and efforts to assure sustained alignment with increasingly changing and often divergent demands? After all, we have called upon the DoD to do everything from assisting with drug interdiction to taking down standing governments. As depicted by the pentagon in Figure 27.1, this raises the fundamental question; what is in and what is out? Often referred to as "framing" we need to first look at what the DoD is currently doing and determine what is in the frame and what should be clearly outside the frame. And equally importantly, is there something currently outside the frame which should in fact be inside. Although these are excellent questions, few have the task of making these decisions at the DoD level. However, this same concept should be applied to all organizations, regardless of size. So how then does one decide what is in and what is out?

27.1 Concept of Value

Before going further a brief discussion about "value" is needed. For the purpose of the following discussions we will look upon "value" as a term denoting a *benefit* derived (yielded) from the combination of resources. The measure of incremental value comparison, or why one value when compared to another is better, occurs when the added value attained from the investment of the same quantity of resources by two different parties yields differing benefits. Also closely linked to this theory is the theory of minimal acceptable value. For example the United States must spend some amount on defense to assure preservation of her Citizens freedom and protection of their way of life—minimal acceptable value. Investments below this amount which yield lesser results would be unacceptable. The challenge comes in determining just what levels of resources are minimally required to produce the acceptable level of "value". Another way of considering this challenge would be to consider that the US spends significantly

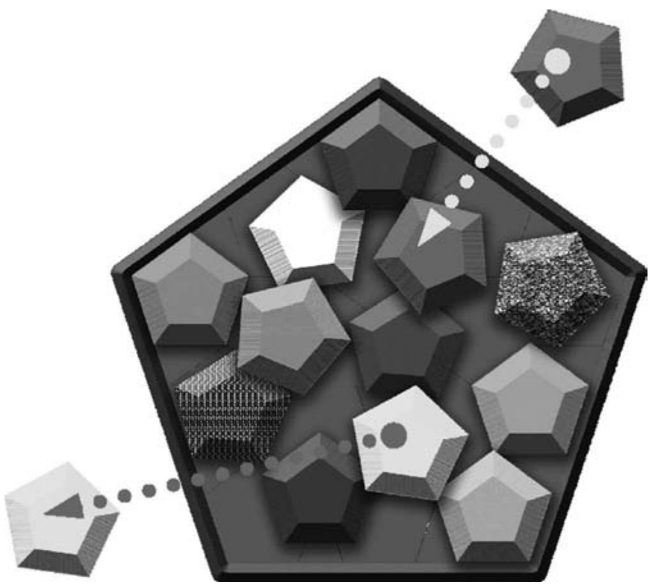


FIGURE 27.1 Pentagon—what is in and what is out?s

more on defense than her next nearest competitor. Does this mean the US is safer? Not necessarily. This is due to many reasons, a few of which include:

- The DoD may be less efficient than its competitor(s) in generating the necessary products/ services.
- The DoD may be producing the "wrong" things with its resources.
- The way the DoD fights may be inherently more expensive. (For example to avoid fighting on US soil the US takes the battle to the opponent thereby significantly increasing logistics costs. Or in order to effectively counter an Improvised Explosive Device or IED the US must expend resources at a rate easily 1000 fold what the enemy spent to put it there.)
- The competitor has methods of defeating US weapons or tactics for a fraction of the cost they take to produce, field and sustain. (For example a hand held laser which blinds the pilot of a billion dollar aircraft and thereby neutralizing the aircraft system.)

Although admittedly the concept of value within the DoD is complex, less tangible than the commercial sector, and more difficult to measure, there is one primary similarity between the DoD and the for-profit business sector—the desire on the part of the consumer that the servicing entity yield "value" at the lowest possible total ownership cost (TOC) within acceptable levels of risk. This need to effectively balance costs with risks is an essential principle which cannot be understated. For an inexpensive product which has a high probability for infant failure will soon find itself without customers. Likewise, "gold plating" a product to the point where the product price far exceeds the benefit derived, will likewise be ill sought.

Additional "traits" which are shared by commercial entities and the DoD and are essential success factors include:

1. Each has the need for technical expertise.
2. Both need to continually "re-shape" their organizations to maximize efficiency.
3. Both must be agile in meeting rapidly changing customer needs.
4. Personnel seek individual stability and growth.
5. Organizational focus must be on the total "value added" (Freeman and Nevins, 1996).

27.2 Why Look to the Commercial Sector?

Where does the DoD find the models, which when employed, assure optimal attainment of organizational strategic aims, thereby successfully delivering the value demanded by the customer? Given the similarities between the DoD and the commercial sector previously discussed, perhaps the best place to begin is by looking to commercial models. But surely we do not expect the DoD to become another General Motors, Toyota, General Electric, AT&T, or Microsoft? While it is true that DoD is indeed unique in many respects, there are also many things about this large organization that are not at all unlike any large commercial conglomerate. All around the world, the United States continues to spread democracy and her economic business models in the belief that this is an excellent model for others to follow. Further, it is commonly believed that prosperous free nation states do not make war against other prosperous free nation states; therefore, this strategy directly supports the United States national security interests.

Business success is centered upon the fundamental principle that the proper employment of resources will yield a product/service which is of greater value than the sum of the resources required to produce it. And further, that this product/service (added value) is filling a customer's needs. Over the years businesses have been perfecting their ability to consistently create and deliver this increased value to customers. In order to do this successfully over the long-term, the very nature of a business entity is one of an organization that is daily fighting for its very survival. A business must continually assess both its external and internal environments to capitalize upon internal strengths while exploiting weaknesses of competitors.* This "business edge" is maintained through the most efficient employment of limited resources in order to yield greater effectiveness than those of its competitors or potential competitors. In short, *commercial business is a war fought on the battlefields of the open market* and supported by a logistics tail that must immediately respond to rapidly changing and fluctuating demands. Its intelligence gathering must be rapid, effective and easily translated into meaningful data from which resource allocation decisions may be made. Accurate assessment of a fluid environment, rapid employment of products/services that were produced below the cost of competitors, the ability to transform with an agility unmatched by others and meet new demands with a speed and efficiency that is unrivaled, is the very nature of business in our globalized economy today.

A for-profit business is usually measured in terms of stockholder equity, return on assets and return on investment. The DoD while not measured in these terms, is measured by the "value" it generates to the American public, specifically, its ability to preserve the peace and her citizens way of life. The DoD is constantly being measured by others in the government, other nations, and most importantly members of the American public.

Given that the DoD and commercial businesses share numerous common characteristics, it is therefore appropriate to consider various commercial business strategic planning principals and tools.

27.3 Strategic Direction and Traditional Barriers to Execution

What is so complicated about strategic direction and getting the job done? After all, shouldn't this be easier in the DoD than in the commercial sector? After all the DoD is designed for leaders to issue orders and for others to follow. So isn't it simply a matter of getting the orders right and issuing them? Figure 27.2 illustrates how this can work, and for many how it should work.

In Figure 27.2, we see that strategy is set at the top of the organization and flowed down. The expectation is that clearly communicating strategic aims throughout the organization will yield the desired results. Unfortunately, time has proven this to not be the case. History is filled with leaders who, while exceptional communicators, failed to meet customer needs. Why is this indeed the case? Unfortunately, the model depicted in Figure 27.2 fails to consider organizational elements and the power associated with individual organizational entities. These entities are referred to as functional organizations as depicted in Figure 27.3.

Over the years, what organizations are called upon to deliver increasingly relies upon the support of more than one functional or system. Without a clear and easy way to horizontally integrate these functions and associated support systems, these organizational structures tend to hinder execution. In these cases the first inclination is for an individual organization to think that if only they directly controlled more of the enterprise assets, then they could solve internal problems and in-turn meet customer needs. Unfortunately, this approach if left unabated, places the enterprise back into the very same position it started from, except that now the enterprise has a single functional attempting to exert disproportionate control over all the rest of the functionals, thereby further exacerbating the problem. Or even worse, the function in a power position assumes control of other functional areas, in the belief that this will solve internal problems. While the organizations that are not likewise expanding feel "under attack" and will subsequently exhibit behaviors that further hinder productive cross-functional workings and in-turn obstruct the development or expansion of integrated systems. This has led to the disparaging designation of functional organizations as "stovepipes". Taken to the next level of thought, the word "stovepipe"

---

* This methodology was first formalized in 1976 by Howard H. Stevenson who is currently the Sarofim-Rock Professor of Business Administration at Harvard University, Graduate School of Business Administration. The Sarofim-Rock Chair was established in 1982 to provide a continuing base for research and teaching in the field of entrepreneurship. Dr. Stevenson was its first incumbent. These concepts are incorporated into a book written by the esteemed business theorist and scholar William F. Glueck, *Business Policy and Strategic Management*, page 186, in the section titled Analyzing Corporate Strengths and Weaknesses Glueck excerpts an essay written by Dr. Stevenson titled "Defining Strengths and Weaknesses," *Sloan Management Review*, 17(3), pp. 51–68.

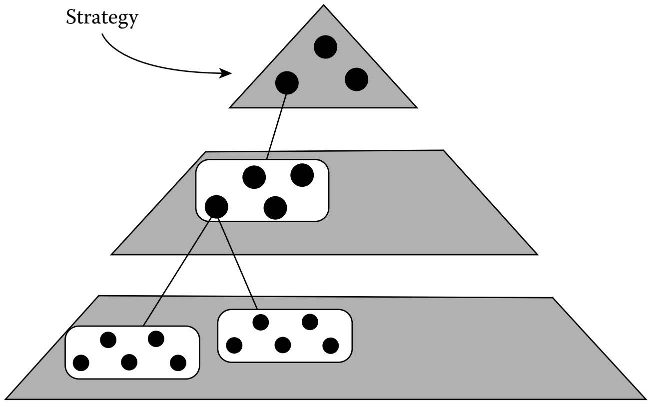


FIGURE 27.2 Traditional view of strategy execution.

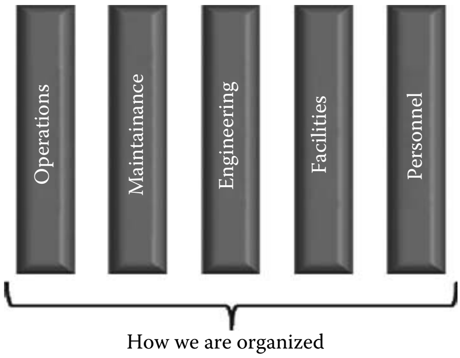


FIGURE 27.3 Functional stovepipes.

indicates that the thickness of the wall of the pipe and its degree of impermeability is directly proportional to the force of the barriers hindering the horizontal integration of the enterprise.

Functional organizations are a critical element of any large enterprise. This is where people learn and hone their skills. They also serve as the mechanism through which individuals achieve increased responsibility, recognition, and promotion—and perhaps most importantly a sense of belonging. Just ask anyone what they "do" and odds are they will respond with who they see themselves to "be". This is especially true the higher you go in an organization above the factory floor. If you ask a person working on an assembly line what they do, you might hear them tell you the end item they assist in producing. For example; cars, refrigerators, computer screens, cell phones, etc. But ask someone above the factory floor and you are more likely to hear; engineer, personnel clerk, production supervisor or maintenance technician. Why is it that the further we get away from the actual production process, the more aligned one becomes with their individual functional stovepipe? And isn't it interesting that the further you are from the production line, the more difficult it is to work cross-functionally. Also of interest is that in Figure 27.3 it is difficult to discern where the customer fits.

This is because the lessons learned on the factory floors seldom, bubble up or rise within an enterprise. Not because people do not want to do well or most effectively and efficiently employ the resources under their control, but rather for failure to look at things from the customer's point of view. When we take the organization and turn it on its side we begin to see some very interesting effects. As illustrated in Figure 27.4 we begin with the customer at the very top. This is who will determine if the proper level of *value* is being delivered. Next we see that Operations is closest to the customer and therefore the best measures

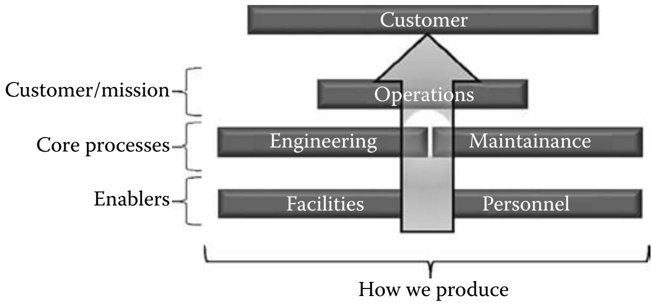


FIGURE 27.4 Aligning the enterprise.

placed upon Operations are those that are set by the customer. Below Operations we see Engineering and Maintenance. This is where many core processes are conducted which yield the internal business outcomes necessary for Operations to deliver. At the very base we see Facilities and Personnel which provide the raw resources necessary to enable the core processes above.

The arrow from bottom to top in Figure 27.4 clearly reflects the *value generation flow* as resources and information combine along this path to produce the product/outcome needed by the customer. We also see a better way to align the organization. This is because the meeting of customer needs is accomplished through the combination of resources, often from various functional areas, which are applied to individual *activities*, that when combined *form a process* which produces *outputs* and *outcomes.* Increasingly, organizations are recognizing the benefit of seeing, and in many cases managing, strategic processes separately from functional areas. When an enterprise chooses to manage a process in this manner, those assigned this task are called *process owners*. Traditionally, the processes being managed by a single person were those that delivered an end item to a customer and cascaded across the entire enterprise, as depicted by the arrow in Figure 27.4. Procter & Gamble pioneered this approach through the establishment of brand managers, where one person for example would be responsible for Tide detergent. In this case, the brand manager for Tide has the goal of selling more Tide, with equal or greater profit margin than was achieved the year before. To accomplish this, the brand manager has visibility across the entire process flow producing these effects, from raw material through consumption and even packaging disposal.

More recently organizations have increasingly recognized the value in additionally defining individual objectives along this path which combine to yield the desired output. In Figure 27.4 we see the impact of looking at the organization differently and can now better discern and differentiate between individual essential process elements by looking through three different lenses; (1) customer (2) internal or "core" processes and those who provide the required resources (mechanisms) who are referred to as (3) enablers. Again, each of these areas; the customer, core process and enablers may have various individual objectives that must be satisfied sequentially from the bottom to the top, to achieve overall success.* Detail objective setting will be discussed later.

- *Customer/mission impact:* The objectives serving this area are targeted to meeting requirements outside the organization. Traditionally, Operations delivers the *effect* produced by military operations which is why they are positioned to interface directly with external customers.
- *Core processes:* Objectives in this area reflect the desired outcomes from various internal processes where resources are combined to produce products/services/capabilities. There are times

* Robert S. Kaplan and David P. Norton, *The Strategy Focused Organization*, Boston MA, Harvard Business School Publishing Corporation, 2001, p79. These commercially available concepts were largely derived from the Balanced Score Card. However, optimal organizational benefits are best obtained when they are combined with the methods and concepts described later in this chapter.

when an individual product or service is provided directly to an external customer, in which case the applicable objective within the core process area would have a customer centric measure. However, this will be the exception.

- *Enablers:* Objectives in this area are primarily focused on people and infrastructure that enable core processes to function, and in-turn satisfy customer needs.

27.4 Where to Begin—Three Pronged Attack

Strategic Process →

Contents and Execution →

Governance →

The preceding concepts are derived from various practices commonly found within the commercial sector. By looking to the lessons learned by others, we now see that the traditional, top down method of deploying and executing strategy is insufficient. We also see the benefit of turning an organization on its side and the need to look at the organization itself in a divergent way discarding traditional methods and seeking new ones which will meet our new construct. Specifically, an enterprise needs to be a horizontally integrated enterprise, which capitalizes on the benefits of functional organizations while simultaneously assuring uninhibited flow between them. And provides for the rapid re-allocation of resources when, and targeted to where, required. Unfortunately, although a good deal of ground has been plowed in this area, previous efforts have focused primarily upon using or adapting commercially available solutions, which have not achieved the desired long-term result. Therefore, we must look to not only the commercial sector, but to combine them with new models and associated practices which will help us better organize and equip or nation's forces. Following is a discussion which will introduce you to a new way of approaching strategic planning and execution which is designed to significantly change how we lead and manage our organizations, yielding unprecedented results at minimal costs.

Let's assume you are tasked to lead an effort with the following direction from the senior leader in your organization. The Commander calls you in and it is clear that things are not going as desired. The Commander states that he has a vision, but is having difficulty seeing it carried out. The Commander is further concerned that, although everyone is very busy, he is not confident that they are working on the "right" things. Your challenge is to take this vision and turn it into executable tasks, that when accomplished across the various parts of the organization, can be readily monitored, resources allocated to where they are needed most, ultimately achieving the leader's vision. To get started you call a meeting with various managers and other leaders throughout the organization. The Commander kicks off the meeting stating his vision and declares that this is priority one, turns the meeting over to you and leaves for another important appointment. Where do you begin?

Imagine you begin by seeking the views of all in the room. After all this would appear to be a reasonable place to start. As people begin to share, inevitably you will begin to hear comments and questions along the lines of: "Where are we really trying to go?" "Why do we need to go there?" "Who decides what we do?" "How will resources be allocated and re-allocated?" "How do we know what we should be doing to best meet the Commanders vision or put another way, the strategic aims of the organization?" "How are we going to do this anyway"? "How is this going to impact my (pet) project that I have invested a great deal of time and effort into?"

All are excellent questions and can be grouped into key areas; the *strategic* planning and execution *process*—"How are we going to do this?" Development of *content* and actual *execution*—"How do we know what we should be doing…?" And *governance*—"Who decides what we do?" Therefore, the first step is to clearly understand these three distinct areas and divide the effort in a way that will address each separately, yet when linked together provide the desired results.

27.4.1 Strategic Process

The strategic process refers to the process which yields the strategic plan and facilitates successful execution. The "process owner" of this process resides traditionally with Plans and Programs within a military organization. As the owner of this process, Plans and Programs is directly responsible for process oversight and successful execution of the planning and execution process. It is critical to note that this organization is *not* responsible for the content of the plan, nor the successful execution of its various elements. It is responsible for providing a process, that when followed will achieve the desired effect, monitoring, and reporting progress in securing strategic aims. This is often accomplished with the assistance of an Integrated Product Team (IPT) made up of members from across the organization. In organizations where other processes do not naturally facilitate cross-functional communication/coordination, this IPT often proves invaluable.

27.4.2 Content and Execution

The "content", which consists of individual objectives, measures, targets and thresholds, are developed to directly support organizational aims. Although the development of content will be discussed in greater detail later, it is important to note that objectives must be developed which, when attained, will meet the needs of the other objectives it supports, and when combined, that these objectives will satisfy customer needs. For an organization to obtain the result(s) demanded by an objective, resources of the organization must be combined into a group of activities, which in-turn form a process. We will be placing individual objectives under the oversight of a single person. To fulfill an objective successfully in an organization that has been turned on its side as depicted in Figure 27.4, the process (or processes) which yield the output/outcome demanded by the objective will cascade across multiple functional entities. Therefore, we now have a single individual who is responsible for a single objective which is supported by a multi-functional process. We call this person an objective process owner (OPO). This OPO is directly responsible for; assuring the content of the objective clearly communicates its intended outcome and that this outcome, when achieved, will meet the needs of other objectives which are dependant upon it. Establishing measures which will accurately track performance, and proposing to leadership associated targets and thresholds for those measures. Additionally, the OPO is responsible for the overall effective operation of the process(s) under their control. To help the OPO, a cross-functional process team may be established. Even if you are in an organization that is not configured this way, every day you contribute to such a process. However, now it will become visible and therefore receive the attention it deserves. Figure 27.5 depicts the various duties of an OPO in descending order.

When viewed in totality, the various objectives required to achieve the desired customer output from an enterprise are grouped into "like" areas. These "like" areas are referred to as goal areas. Goal areas consist of multiple objectives and lend themselves to being managed by a single senior leader within the organization. An example of a goal area may be a grouping of objectives within the *enablers*' strata that is titled Enabling infrastructure. Under this goal area would fall everything that is needed to support the physical and information needs of those working within the enterprise. Such things would include; computers, office furniture, equipment, buildings and financial data visibility. You will note that these items fall within separate functional organizations. However, the individual responsible for this overall goal area is responsible to assure a *capability* is available to meet the needs to *enable* various organizational processes and in-turn organizational objectives. Because this individual has the sole responsibility for this capability area, they must also be given the authority to examine and when necessary alter underling processes. This is the single person to whom the organization will turn for this capability and therefore this person is referred to as a "Champion". A Champion, is the individual who has authority over the process or processes required to deliver a defined capability, and therefore has the final authority for all underlying process designs. Note that the Champion will not "own" all the resources required to execute the process or processes for which they are responsible. However, they are responsible to assure

FIGURE 27.5 Objective process owner (OPO) responsibilities.

that the processes within their goal area optimally achieve the desired output/outcome. Champions are normally senior officers or civilians. Each Champion is responsible for directing and coordinating actions (generally executed at the objective level) to achieve the strategic intent of the assigned area, and reporting performance to senior leadership. Champions work closely with OPOs, frequently reviewing performance and progress toward closing performance gaps. Champions assist with implementation of improvement initiatives, and recommend changes to strategic objectives when appropriate.

Specific responsibilities include:

- Direct activities associated with accomplishing the overall area of responsibility, including actions that fall within the responsibility of multiple functional organizations.
- Oversee strategic objectives that fall under their area.
- Approve performance metrics developed by OPOs for each objective within the area.
- Brief area performance by individual objective to leadership, and when appropriate, recommended actions through the Board and in-turn the Council (see discussions on Governance).
- Consolidate findings and conclusions regarding progress on the strategic objectives under each area.
- Review other areas and associated strategic objectives to discern linkages and, where appropriate, influence external: measure, target and threshold setting.
- Advocate for resources required to achieve acceptable performance.
- Recommend changes to the Strategic Plan (e.g. deleting, adding, or changing a strategic objective) when appropriate.
- Call joint Champion meetings as required.
- Attend joint Champion meetings when initiated by another Champion or Plans and Programs.

27.4.3 Governance

Leadership is responsible for providing the strategic direction and overseeing and approving the entire process. Leadership plays an active role in the deliberate planning and execution process which takes the organization from Vision to executable tasks. Most importantly, this structured engagement affords leadership unprecedented visibility and inextricably links the resource allocation process to the strategic execution process. The structure within an enterprise that enables these actions and associated outcomes is referred to as "Governance" which is discussed in greater detail in the next section.

Strategic element	Components	Flight lead	Status
Strategic process	• Develop strategic process for planning and execution • Process oversight	• A8 (plans and programs) • Strategic planning and execution IPT	• Process complete • Work begun on IT feed
Content and execution	• Objectives • Performance measures • Targets and thresholds	• Champions • Objective process owners	• Phase 1 - Complete • Phase 2 - Measures complete - Target and threshold setting in work
Governance	• Strategic direction • Construct approval • Content approval • Oversight • Resource allocation	• Commander/ organizational leader (and staff)	• Vision complete • Strategy planning and execution roll out Jul xx

FIGURE 27.6 Three thrust areas of strategic plan development and execution.

Figure 27.6 depicts the three prongs of attack; Strategic Process, Content and Execution and Governance. It lists the associated components of each and provides a column to list the individual or organization responsible for the associated component action. The last column is used for status reporting. This is an excellent tool for observing the progress of the entire effort of building a plan and implementing a sound execution process.

27.5 Governance—How it Works

Governance is a critical element in assuring successful attainment of vision to task and subsequent satisfaction of customer needs. This is because of one thing—money. The governance structure defines the requirements and allocates resources. To accomplish this, many defense organizations have a process in place which reviews and approves budgets and the distribution of financial resources. In Figure 27.7 we see on the bottom right an entity titled Command Group which feeds into the Board which feeds into the Council. This is the traditional process. To that structure you will see the addition of the various entities described under "Where to begin—three pronged attack". Note the various relationships and associated actions performed by each.

While there are many ways to establish a sensing responsive governance structure, the one depicted is an example of a structure which is proving successful. In this new governance construct the Command Council acts as a deliberative body with responsibility for monitoring the overall attainment of the Vision, and associated supporting strategic objectives. The Council would meet as specified by the Commander to review progress, allocate resources and adjust strategic direction as required. The principle goal is to construct a governance structure which will monitor strategic development to assure alignment with the aims of the enterprise and rapidly move resources to where they are needed. Exactly how this works will become clearer following examination of the 6-Phase Strategy Development and Execution process discussed in the next section.

Specific responsibilities include:

- Meet as directed to confirm strategic direction.
- Modify strategic objectives to meet changing/emerging requirements.

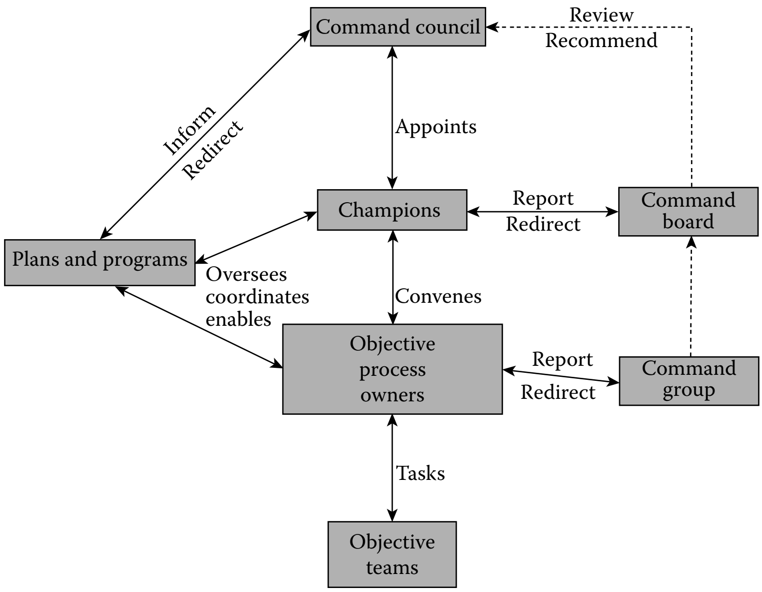


FIGURE 27.7 Typical DoD governance structure linked to strategic execution.

- Monitor progress towards achievement of grouped and individual strategic objectives.
- Allocate resources to address strategic gaps.

27.6 The Strategic Process

With a basic understanding of governance and the roles and responsibilities of Objective Process Owners and Champions, it is appropriate to turn our attention the process which is used to build strategic plans and execute them in a way that links vision to task. What is needed is a process that will assure successful vision attainment, with the expenditure of minimal resources. The key is to put into place a system which provides adequate oversight, facilitates creative thought, assures independent actions complement each other, and allow for near real-time resource allocation adjustments as needed. The 6-Phase Strategy Development and Execution Process Model is designed to serve as the overarching process model which successfully accomplishes this by translating vision into executable monitored tasks.

Strategic planning and execution is all about translating vision into executable tasks, assuring resources are best aligned to support these required tasks, and simultaneously providing a simple way for leadership to monitor performance. To achieve these goals the Commanders' Vision must be linked to various specific strategic objectives which are measured. These objectives drive desired actions (tasks). The result of these efforts yields outputs in the form of products/services/outcomes which, when measured against associated targets and thresholds, define performance. Performance levels other than those desired represent "gaps". Leadership engagement when these performance issues become apparent is critical to assess the gap and direct action which may result in the requirement to expend additional resources.

The 6-Phase Strategy Development and Execution Process model depicted in Figure 27.8 is a disciplined approach which was developed to close the gap between various other models in use throughout the commercial and government sectors. Using fundamental Systems Engineering (SE) practices and principles, this model focuses on the processes and effectively addresses the interstitial spaces between vision setting, strategic plan development, execution, data gathering, data analysis, reporting and

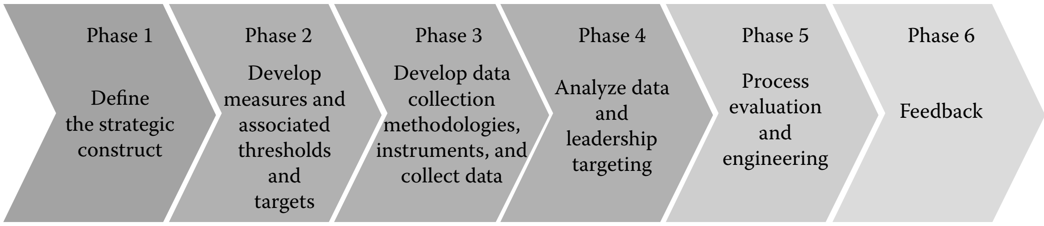


FIGURE 27.8 Freeman 6-phase strategy development and execution process model.

governance. The model assures that strategic elements of the planning and execution process are performed in the order required. It further defines the office and/or individual responsible for this activity and clearly denotes where to incorporate process evaluation and improvement (Phase 5) thereby linking process improvement efforts directly to strategic actions. The model relies upon the construct that processes, not individuals or organizations, yield products, services and outcomes. Part of this model includes the recognition and assignment of OPO's and Champions, discussed earlier, who are responsible for process outputs. Finally, although depicted linearly, this is a looped continuous process as represented by Phase 6.

To gain a better understanding, we will now beak down each of the six phases of the model depicted in Figure 27.8. While examining each of these phases individually here are some things to keep in mind.

- Although drawn horizontally, this is a circular process.
- For the model to function at its best, feedback must occur throughout the entire process, uninhabited by any perceived phase or organizational boundaries.
- An enterprise using this model will find it significantly beneficial to form a cross-functional IPT/working group whose members are tasked to develop the plan and initiate its execution. This will allow time for OPO's and Champion's, who are assigned as part of this 6-Phase process, to get up to speed and eventually take the reins from the cross-functional working group.
- Each Phase of the 6-Phase model must be accomplished in sequence. Failure to follow the Phases in order will yield unsatisfactory results and also drive undesired organizational behavior. The larger the enterprise, the greater the probability that more than one organization will be involved for each Phase. Further, multiple organizations may be involved with individual Phases. Therefore, if a single organization follows/yields to individual organizational needs, over those of the entire enterprise, there will be temptation to execute Phases out of sequence. Imagine for example, that you are responsible for improving processes within your enterprise. (Phase 5) You are receiving a great deal of pressure to show savings through process improvement initiatives and look for a place to begin. However, the enterprise has not yet completed development of measures (Phase 2). How will you know what processes are strategically significant to the enterprise? How will you know which of these are underperforming and therefore require improvement? The reality is that you will not. Therefore the best you can hope to do is to go after the "low hanging fruit". But even then, you run the risk of consuming treasured enterprise resources to improve a process that has not yet been defined in strategic terms and therefore may not have strategic significance. Further, you may force the definition of a process without a clear understanding of customer needs and linkage to the strategic aims of the enterprise in meeting those needs. Failure to link process improvement efforts to the strategic aims of the enterprise is like firing a weapon without calibration or aiming. It may hit something, but more often than not it will do more damage than achieve its desired effect.
- Once through Phase 2, individual objectives actions associated with this objective will proceed through the model at different rates.

- Underperforming objectives should attract resources to bring performance to the desired level—above the threshold. Over performing objectives—those performing above the target should be examined for reduction of resources (Target and Threshold setting will be further discussed later.)
- Proactive senior leadership (those to whom the Champions report) engagement throughout the entire 6-Phases is essential, especially at critical decision points which are depicted in Figure 27.9 and will be discussed in greater detail under each Phase.

27.7 Phase 1: Define the Strategic Construct

In Phase 1 the enterprise will:

- Clarify its vision
- Define strategic principles
- Define perspectives
- Define goals
- Define objectives

Phase 1 is centered upon leadership's proactive engagement, before the expenditure of any resources. All too often those working within an enterprise will invest heavily to improve a process only to learn later that their efforts, while well intentioned, missed the mark. This is the fastest way to poison the process improvement well from which others will attempt to draw in the future. The process outlined in these six phases is a full contact activity and all players must be on the field, most importantly—leadership.

An enterprise that remains stagnant will eventually become irrelevant to its customers and evaporate. Therefore, it must be continually moving in a direction that meets both current and anticipated future customer needs. The leader of a successful enterprise must have *vision* and be able to articulate that vision throughout the enterprise. Once defined, the leader's task is to unbridle members of the enterprise, as they are best suited to figure out the "how". But without a clear mutual understanding of the various essential elements outlined in Phase 1, the leader may observe the expenditure of treasured enterprise resources that do not produce the desired effect.

What does visioning look like? Two excellent examples of leadership visioning on a grand scale include; President John F. Kennedy's 1961 Urgent Needs speech to Congress in which he stated;

> "I believe that this nation should commit itself to achieving the goal, before this decade is out, of landing a man on the Moon and returning him safely to the Earth. …This decision demands a major national commitment of scientific and technical manpower, material and facilities, and the possibility of their diversion from other important activities where they are already thinly spread. It means a degree of dedication, organization and discipline which have not always

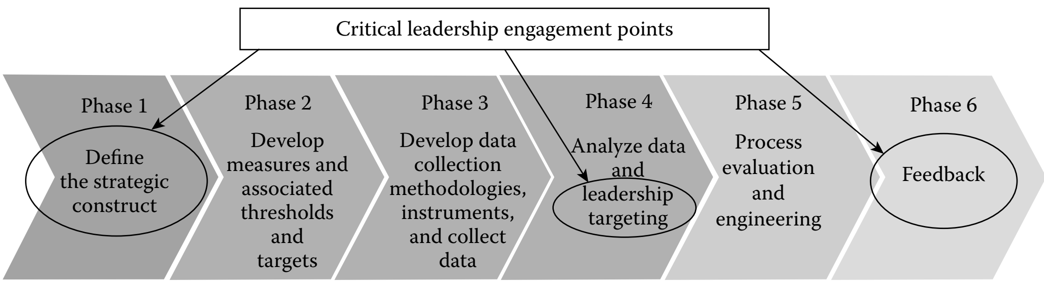


FIGURE 27.9 Critical points for leadership engagement.

> characterized our research and development efforts. It means we cannot afford undue work stoppages, inflated costs of material or talent, wasteful interagency rivalries, or a high turnover of key personnel."*

The second example was in the early 1980s when Jack Welsh, shortly after assuming the position of CEO at General Electric (GE), told the leaders inside GE that their business units had to be number 1 or 2 in their business within 5 years, or GE would no longer be in that business. The vision here is clear, be the big boy on the block. For Jack Welsh believed that those with the greatest market share were best positioned to control that market, while simultaneously capitalizing on economies of scale that could not be matched by the competition, thereby increasing flexibility and market maneuverability and in-turn survivability within that domain.

The *strategic principles* illustrated here were to; identify the business you are running for GE, identify your competitors and compare yourself to them. Once known, a business unit leader achieved success by being no less than number 2 in that market. Also, as with President Kennedy's speech, there is a clear recognition of the associated financial aspects, yet another fundamental strategic principle.

Therefore, *strategic principles* serve to define the playing field by establishing broad expectations related to the vision. President Kennedy did this when he set the vision of going to the moon, but then added the strategic caveats that dealt with the effort required and need to change.

Once the vision and strategic principles have been established, we turn to *perspectives*. Perspective setting helps us to put the vision into context. President Kennedy declared his moon challenge in the context of the competing interests between the United States and the United Soviet Socialist Republic when in the same speech he stated;

> "… if we are to win the battle that is now going on around the world between freedom and tyranny, the dramatic achievements in space which occurred in recent weeks should have made clear to us all, as did the Sputnik in 1957, the impact of this adventure on the minds of men everywhere…".

The clarification of strategic perspectives in association with the vision and strategic principles helps throughout the 6-Phase process. These three combine to create a clearer understanding of what is being sought, the process that will be used to get there and the environment within which these events will occur. They serve as a central rallying point and help to keep members of the enterprise focused on what is important over that which is less important or trivial. They also help to focus objective development and later in Phase 5 with alternatives analysis. This increased understanding—at the outset—builds flexibility into the decision making process. Armed with this shared knowledge, decision makers have a better understanding of the underlying intent and therefore it is easier to make decisions when data may not be readily available or as complete as desired.

Next, the task at hand may be large enough to warrant the establishment of overarching *goals*. In the 6-Phase environment a goal defines an area that is comprised of two or more objectives. For example the overarching goal may be to assure timely availability of mission capable resources. To achieve overall success, various individual objectives dealing with training and equipment availability may be required. Note that goal areas may not always be required. As discussed earlier, if a goal area is established the single individual overseeing goal attainment is called the Champion.

Once the overarching goal areas are defined individual objectives are constructed which, when met, will assure the goals are met. Depending upon the task at hand, it may be equally acceptable to begin building individual objectives first and then, if desired, group various objectives into goal areas.

---

* Special Message to the Congress on Urgent National Needs, President John F. Kennedy, May 25, 1961, Presidential Files, John F. Kennedy Library, Boston, MA, http://www.jfklibrary.org/Historical+Resources/Archives/Reference+Desk/Speeches/JFK/003POF03NationalNeeds05251961.htm.

Therefore, objectives may either stand alone or be grouped. Recall that individual objectives are managed by Objective Process Owner's (OPO's). When building an objective it is essential to not be hindered by the method or way this objective will be measured. Measurement is intentionally removed from the activity of building objectives to avoid hindering accurate development.

When leaders are given the responsibility to assure the successful achievement of strategically linked objectives, verses organizational metrics, leadership focus shifts from stove-pipes to a horizontally integrated–process-based enterprise.

27.8 Phase 2: Develop Measures

In Phase 2 the enterprise will:

- Define measures
- Establish thresholds
- Establish targets
- Define data characteristics

Measures are a critical element of any strategic planning and execution process. They guide action and provide senior leaders with easy-to-use indicators of progress in the attainment of assigned objectives.

Objective Process Owners use measures in conjunction with the associated objective to develop "thresholds" and "targets". When applied to an individual measure, a "threshold" reflects the minimally acceptable level of performance while the "target" reflects the optimal desired level of performance. Champions and OPOs' should describe performance using thresholds and targets.

It is anticipated that measures will continue to be altered/refined as the enterprise proceeds through the 6-Phase process and/or areas of strategic emphasis change. However, the most fluid elements of the measurement process are threshold and target descriptions. While objectives serve as the vector heading, targets and thresholds become the "throttle". Meaning that to "push up" performance; a threshold may be increased, for example from 65 to 80%, thus demanding a higher level of performance and in-turn greater resources to achieve. The inverse would also apply. Therefore, if an organization desires to publish a formal strategic plan, given their purpose and intended fluidity, targets and thresholds should not be part of a published plan.

The last part of Phase 2 deals with defining data characteristics. With the desired measure established and associated thresholds and targets set, it is now time to turn to the data portion of this process. Up until now, we have intentionally refrained from addressing data requirements and the associated IT requirements. This was intentional as all too often an enterprise if falsely bound by its IT infrastructure, finding itself at the mercy of legacy systems and associated imbedded business practices—practices which were customarily developed by a programmer who did not have visibility or understanding of overarching enterprise requirements. Armed with the information developed in Phases 1 and 2, the system architect is now able to gain a better understanding of the requirements and in-turn is significantly better positioned to design solutions which will be aligned to best satisfy enterprise strategic needs. It is recommended that the Department of Defense Architecture Framework be used to guide systems architecture development.

When defining the data characteristics associated with gathering the data needed to provide the information demanded by a measure, the designer must often consider the fact that it may not be reasonable to gather 100% actual data. In these cases the designer must consider the characteristics necessary for the data which is available. For example can the population being examined be divided into sections (strata) that are internally homogenous? This is called stratification. Is it best to begin by sampling various clusters within a stratum? This is called clustering. Sampling weights also need to be considered as does the population number of the sampling effort. These and many more, characteristics of data are considered and initial applicable algorithms developed at this stage of the 6-Phase process. In addition to the data characteristics, the data architect must assist in identifying the source of the data required. Often, various OPO's will require the same or similar data. Therefore, the goal is to identify and/or establish a single source from which to secure "like" data. This source must provide consistently accurate (authoritative) data which can then be shared. When this is accomplished successfully, it is refereed to as *shared authoritative data.* Often this single source was previously providing data to a single functional system, and when the use of this data is expanded the owner of that system must be aware of the data's new use, and coordinate any future changes which will affect this data with a wider audience. With this effort to seek shared authoritative data and use that single data source as needed across the enterprise, to meet process rather than organizational needs, we can see the foundational building blocks of achieving a horizontally integrated enterprise. For these reasons, it is essential that the data architect have open and frequent communications with all customers (OPO's) throughout the entire data designing process.

27.9 Phase 3: Data Collection and Analysis

In Phase 3 the enterprise will:

- Develop data gathering methods
- Collect data
- Define data presentation formats

In Phase 3 the system architect will continue development of various views of the enterprise architecture. As recommended in Phase 2, it is recommended that the Department of Defense Architecture Framework continue to be used to guide further architecture development. Once defined the system architecture will be used to guide the execution of data collection.

There is no shortage of data, and therefore collection must be targeted to gathering only the data required. OPO's must be confident that the data being gathered is accurate and can be riled upon for decision making. Further, it may be necessary to interpret the data and this too is an area where the OPO must be fully engaged. Assumptions and algorithms must be equally scrutinized and understood.

The OPO and/or Goal Champion must clearly understand the way in which data is presented for analysis. The purpose is to assure that the data is collected in a way that optimizes its later presentation for analysis. For example, if spreadsheets will be used for analysis, then efforts should be made to have the data sourced in a format that is readily digestible by this method. Another example would be data which is derived from legacy systems and may arrive with hidden assumptions or algorithms which must be accounted for.

The goal of Phase 3 is to secure only the data needed, in a manner that consumes minimal resources and arrives in a form that is readily analyzed.

27.10 Phase 4: Data Analysis and Leadership Targeting

In Phase 4 the enterprise will:

- Develop analysis methodologies
- Analyze the data
- Present results

Phase 4 is the data analysis and presentation phase. Analysis methodologies designs begun in Phase 3 are finalized in the beginning of Phase 4. Here again it is essential that the OPO and when appropriate the Goal Champion, be proactively involved in the process of defining analysis methodologies. While they may not be the ones to actually analyze the data, a fundamental understanding of this is essential. Additionally, the OPO/Goal Champion must assure that the ultimate result will accurately reflect/describe the status of objective attainment.

Once the analysis methodologies have been established analysis is performed. The OPO guides this analysis with the goal of securing information which will best describe the current state. The OPO then guides the format this information will be presented to the Goal Champion and leadership through the governance process. Every effort should be made to use commercial "off the shelf" software for tracking and reporting.

With a firm understanding of the current state of performance, leadership is able to put one objectives performance in context of all other objectives and associated goal areas. This then allows leadership to best target the limited resources of the enterprise to those processes in need of improvement and deemed strategically significant. It should also be noted that objectives found to be over performing are candidates for resource reduction.

Therefore, the goal of Phase 4 is the balancing of resources and accurate targeting of process engineering/reengineering initiatives undertaken in Phase 5.

27.11 Phase 5: Process Evaluation and Engineering

Leadership now has the information needed to proactively engage and make a decision. If the objective is being satisfied that decision may be to do nothing. Likewise if the objective is not being met, leadership may choose to accept that level of performance, or direct that action be taken. Note the proactive decision making required by leadership at the outset of Phase 5. Let's assume that an objective is not being met and leadership has directed that action be taken.

To meet an objectives threshold/target we must first seek to understand what is producing the current state. This points us to the underlying process or processes which are yielding the undesirable result. Therefore, we have now been "pointed" to the process or processes in play. This is the first time we are formally examining a process for potential alteration/improvement based upon leadership direction which resulted from an observance of an out of compliant strategic objective. Hence we have successfully linked the targeting of process improvement to our strategic aims.

There are numerous process improvement techniques available. However, each is customarily designed and/or grew out of a particular functional area of the enterprise. One excellent example of this is the Toyota Production System (TPS) which makes wide use of lean manufacturing techniques. The tools contained in TPS which include Rapid Improvement Events (RIE's) and Value Stream Mapping (VSM) are excellent when applied correctly. Six-Sigma, originally developed by Motorola and used widely throughout General Electric, is another example of an excellent tool which was designed primarily for the manufacturing arena. As with any tool, they should be used for the application for which they were designed. The TPS tools and Six-Sigma grew from the factory floors and therefore, yield the best results when applied to linear processes which can be readily broken into relatively simplified segments. Attempting to use these tools at higher levels within the enterprise is often problematic for the same reasons we discussed earlier when examining organizational structures.

Tools that are designed for use above the factory floors are also available and fall primarily under the heading of enterprise reengineering. These tools are also useful, especially when an enterprise has the flexibility to alter targeted customer sets and/or the products or services produced. Unfortunately, this is often not the case for the DoD. Therefore, the members of DoD must be fully cognizant of various tools, their individual capabilities and assure their use is targeted to best address the challenge at hand.

From this brief discussion you can see that different tools are used to meet different needs. At this point it is important to recognize that one size does not fit all, although various developers and proponents of individual process improvement tools and grouped toll sets often become anchored in that belief. Therefore, rather than focusing on a specific tool or method to solve DoD enterprise-wide challenges, the Six-Step Process Change Model was developed and is depicted in Figure 27.10. This model operates within Phase 5 of the 6-Phase model and is applied to individual processes under examination. As with the overarching 6-Phase Strategy Development and Execution Process Model previously presented, these steps must be accomplished sequentially. You will note that Step 1, "identify opportunity"

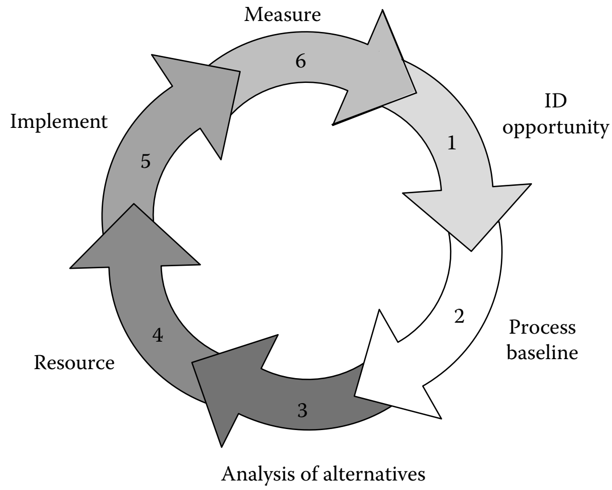


FIGURE 27.10 Six step process change model.

was completed as a byproduct of accomplishing Phases 1–4 of the 6-Phase model and therefore Phase 5 serves as the entry point to Step 1 below.

27.12 Phase 6: Feedback

In Phase 6 the enterprise will:

- Measure results (actual to planned)
- Feedback to leadership
- Course correct

Phase 6 represents the feedback loop that all self correcting systems must have for stability. Successful attainment of strategic aims demands a system which is agile and can rapidly realign resources as new information becomes available. Although depicted as a separate Phase, in fact this process is designed to facilitate as near to real-time feedback as possible. For example, we have seen where the senior leader(s) is fully engaged in Phase 1, providing "feedback" to the enterprise, fully engaged again at the beginning of Phase 5, and Phase 6 assures timely formal structured feedback. This illustrates the inherent flexibility of the 6-Phase process, allowing for continuous revision and improvements as requirements shift, funding changes, or the enterprise chooses to accept more risk in a resource constrained environment. Here the linkage to the formal corporate structure as discussed under the Governance section is an equally essential element of this process. Figure 27.11 illustrates a typical operations tempo for structured reviews.

27.13 Summary

In this chapter, we have presented various theories, and introduced a model which aligns an enterprise to optimally meet customer requirements. We have learned a new way to look at the enterprise and how to organize around processes which produce the outputs/outcomes needed by our customers. The 6-Phase strategy development and execution process provides a disciplined approach which successfully transforms an enterprise into an environment that fully aligns and invests its limited

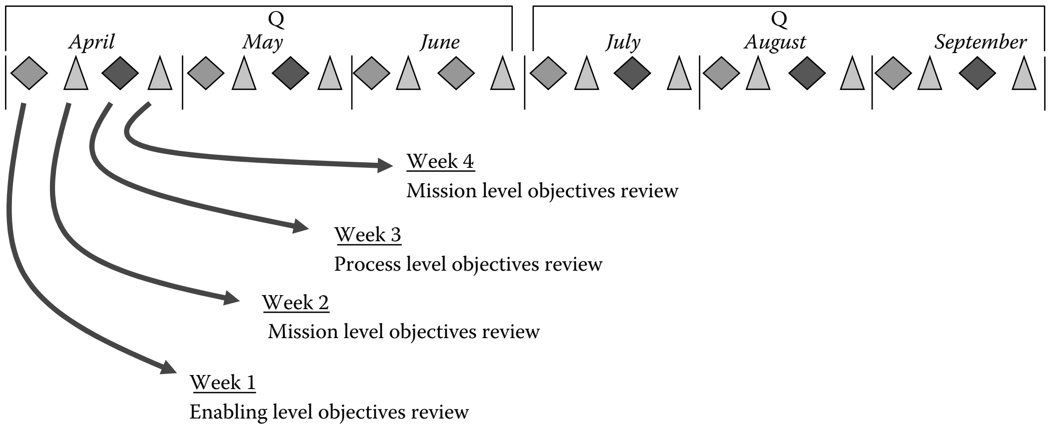


FIGURE 27.11 Typical review schedule.

resources targeted to achieving the mutually agreed upon strategic aims of leadership. Armed with this information you can now take an enterprise beyond historical stovepipes and simultaneously successfully link strategic aims to measurable tasks, in a dynamic highly responsive and rapidly adaptive environment.

Bibliography

Abell, D.F. and Hammond, J.S. 1979. *Strategic Market Planning, Problems and Analytic Approaches*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Executive Office of the President Office of Management and Budget. 1999. *Federal Financial Management Status Report and Five Year Plan, Executive Summary*. Washington, DC.

Freeman, Col G.., and Nevins, Capt L. 1996. *Technical Atrophy's Threat to Large Organizations*. AF Aeronautical Systems Center Press, ASC-TR-96-5003.

Garrison, R.H. and Noreen, E.W. 1997. Managerial Accounting. Boston, MA: Irwin McGraw-Hill.

Glueck, W.F. 1980. *Business Policy and Strategic Management*, 3rd Edition. New York, NY: McGraw-Hill Book Company.

Grossman, E.M. 2002. *New Directive Requires Defense Agencies To Track Their Readiness*. Washington, DC: Inside the Pentagon, July 4.

Hummel, R.P. 1987. *The Bureaucratic Experience*, 3rd Edition. New York, NY: St. Martins Press.

Kaplan, R.S., and Norton, D.P. 2001. *The Strategy Focused Organization*. Boston, MA: Harvard Business School Publishing Corporation.

Kinnard, D., 1977. *The War Managers*. Hanover, NH: University Press of New England.

Lawrence, P., and Littman, I. 2001. *Innovation*. Washington, DC: The Price Waterhouse Coopers Endowment for The Business of Government.

Liberman, J., Senator. 2002. The Best Defense: Leveraging the Strength of Our Military to Protect the Homeland. http://www.senate.gov/-liberman/speeches/02/06/2002626836.html, accessed June 26, 2002.

Maier, M.W. and Rechtin, E. 2002. *The Art of Systems Architecting*, 2nd Edition. Boca Raton, FL: CRC Press.

Mercer, J. 2002. *Performance Budgeting for Federal Agencies, A Framework*. International Business and Technology Consultants.

Naval War College. 2000. *Strategy and Force Planning*, 3rd Edition. Newport, RI: Naval War College Press.

Niven, P.R. 2003. *Balanced Scorecard, Step-By-Step for Government and NonProfit Agencies*. Hoboken, NJ: John Wiley & Sons.

Office of Management and Budget, Circular A-11. 2001. Overview of Strategic Plans, and Annual Program Performance Reports, Washington, DC.

Rainey, H.G. 1991. *Understanding and Managing Public Organizations*. San Francisco, CA: Jossey-Bass Publishers.

United States Department of Defense. 2007. DoD Architecture Framework, Version 1.5.

United States General Accounting Office. 2000a. Report to the Ranking Minority Member, Committee on Governmental Affairs, US Senate. Performance Reporting, Managing for Results, Challenges Agencies Face in Producing Credible Performance Information.

United States General Accounting Office. 2000b. Report to the Ranking Minority Member, Committee on Governmental Affairs, US Senate. Performance Reporting, Managing for Results, Emerging Benefits From Selected Agencies' Use of Performance Agreements.

United States General Accounting Office. 2001. Report to the Ranking Minority Member, Committee on Governmental Affairs, US Senate, Department of Defense. Status of Achieving Outcomes and Address Major Management Challenges.

United States General Accounting Office. 2002. Report to the Ranking Minority Member, Committee on Governmental Affairs, US Senate. Performance Reporting, Few Agencies Reported on the Completeness and Reliability of Performance Data.

United States Naval War College, Syllabus for National Security Decision Making, Strategy and Force Planning and National Security Decision Making Planning Guides, 12 August–8 November 2002.

Weinstock, M. 2002. *Managing Results, Perfecting Performance*. Washington, DC: The Government Executive.

28

The Military Performance Team

Jesse W. Brogan
Walter Reed Army Institute of Research

28.1 Team

The source for this term is in the use of dumb animals. "Team" was first used to address horses or oxen that were yoked together and used for accomplishing work. This team did nothing on its own. Someone had to select the animals and fit them into the yoke. This team required a driver to give it purpose and direction, and to apply incentive to work.

We find "team" applied to different effect when it came to athletics. There was a group of individuals in competition with another group. There was a great difference from the animal team as people were intelligent and self-motivating. They did not have to be selected and yoked, but simply brought together and given a common purpose. The athletic team members provided their own internal incentives.

At this point, we can identify two general types of intelligent teams. We have a number of children who gather together for a game of football; compared to a structured effort by adults who are specifically selected and managed for a performance effect. The "common team", that of the youngsters, is self-managing and its members choose internally which person will do what specific activity. The performance team, exampled by the professional athletic organization, has leadership that makes decisions affecting who is on the team, and what each team member will do.

The common team members do what they are able. The performance team members do what they are directed to do by those who are in charge. The common team members bring their natural abilities and motivation for use in a team effort. The performance team member receives special training to improve what they do; they practice designed plays, and they practice in the team roles they will fill developing personal expertise and learning how to perform most effectively with one another.

Common teams do not compete successfully against performance teams. The performance teams are far more effective, even beyond the obvious advantage of having more able members.

28.2 Teaming in Business

In the history of business, we find the same sort of split in team concepts.

Going back to the late 1800s, we find performance being accomplished by work gangs; and these were in the charge of bosses. The gang was a collection of workers, hired to do the work with what they would bring to work with them. The boss was there as the agent of the owner; the boss was someone who would direct the gang to a task, keep them working until the task was done, and would receive the result on behalf of the business. The gang was a common team where those who were on the team divided up the task as seemed most fair and appropriate in their understanding.

Frederick Taylor, in his 1911 work on scientific management, demonstrated the benefit of the performance team. He promoted selection of the best workers for the type of work being performed. He studied the performance processes, and determined a best way to gain the result. Then he gave each worker specific training in that best performance technique. He provided the workers on this team with best tools, best processes, and a manager who was there to do what was necessary or convenient to assure the performance result accomplished by the other team members.

As should have been expected, there was no real competition. Taylor's managed work group out-produced the boss-led gang efforts by three to four times. The difference was intelligent management.

28.2.1 When People are Managed for Performance, they Out-perform those who just Gather for Performance

So what makes the performance team so much more effective?

The answer is not simply in processes. The answer is in management; and it addresses attitude and approach. Those on the performance team may not want to succeed any more desperately than those on the common team; but they have a structured effort that is maintained by someone who is dedicated to a managing and supporting role. They are intentionally managed; in contrast to the common team that is self-managed by its working members.

The performance-based athletic team is lead by a performance-based coach. If the team loses; the coach also loses. The performance-based leader is someone who shares with the team; and the success of the team will often be the direct result of what the coach does in organizing and preparing team members to take part in the larger effort.

The same was true in the performance initiated by Taylor. The foreman and worker shared in the success or failure of group performance, for they both had their part to play in assuring any success. If the worker failed, it was also a failure for the foreman; their performance responsibilities were based on the same end result. They had to team to make it work.

There is one important characteristic to be noted. The performance team does not result from the gathering of people for a purpose. The performance team requires an intentional effort that involves application of intelligence; and it takes a positive application of leadership to bring the performance team into existence.

Without this application, we find the common team; which is far less effective at gaining results through the efforts put forth by team members.

28.3 Military and Teaming

Soldiers are honed into smaller operating units that function as performance teams.

Performance teams are the organizational answer to a need for performance, as people working together in intelligent and directed-coordination always out-perform common teams.

There are intentional efforts to establish performance teams as the lowest working element of any modern military organization. Performance teams survive and succeed; common teams often fail.

There is an officer in immediate charge of a performance team. The officer is a central performer, one who manages and maintains the team, and who shares in the successes and failures of the team.

Most remarkably, we look to more senior officers, and we usually find common teams, where some senior officer directs subordinate activities to a number of juniors, trusting that their training will allow them to work effectively together to assure a central accomplishment. It is obvious that the performance lesson has not been learned above soldier or sailor level.

The nonperformance attitude is startlingly obvious. If an officer is lost, he or she can be readily replaced by another officer who has the same basic education and training. Instead of maximizing the effectiveness of the officers who direct soldier performance teams, the military organization has multiplied the number of officers. The modern military has attempted to make up for lack of performance quality by increasing the quantity of officers. Our modern military leader has not been prepared to enter into effective performance teaming with those who will implement the officer's orders and directives. As a result, a substantial level of improvement is possible in officer-team performances.

Even as a soldier-level team best assures a performance result from the soldiers, the performance team at the officer level will best assure a larger result through those soldier-team units. This is not new knowledge, but a new application of what military officers already know.

Lacking performance teaming in the officer corps, we find the same lack of teaming with senior civilian employees. They are given directions; and they are held to account. The modern military officer is trained and encouraged to be a leader, but to abstain from being a real part of the performances they lead. The result is common teams, individual subordinates gathered to work with what they bring, and otherwise left to coordinate their own efforts, to divide up the work as seems best in their local understanding, and to establish subordinates common-teams to do the work.

Military officers, who would not send a common team of soldiers into harms way, readily send common teams of subordinate officers, of civilian leaders, and of civilian employees to expend their military dollars on common-team efforts.

28.4 Establishing the Performance Team

Management is the key to establishing a performance team. The common worker will not be able to team with a leader unless the leader will team with the worker. The athlete cannot team with the coach unless the coach is willing to team with his athletes. The senior officer will only establish a performance team when that leader is willing to team with his subordinates.

This isn't rocket science; it is just cross purpose to modern military thinking. It starts with the performance-team leader accepting a personal responsibility for the success of every person on the team. It starts with the acceptance in this leader that he or she will do whatever is necessary and convenient to assure each subordinate success. It starts with the performance leader becoming the center of the performance.

28.4.1 Performance is not Limited to what gets Assigned to Subordinates. Orders cannot Create the Performance Team

We must start at the beginning. Performance can only be gained where there is a clear concept of product, something that has to be gained through the team effort. Without this, no team can be developed to perform it. It is the performance leader who must start with the understanding of product, and must accept personal responsibility for the management effort that will gain that performance through others.

28.4.2 The First Key is Product, Not Command!

The mechanism for establishing a performance team is not difficult, or even unexpected. It is up to the leader to determine subordinate performance requirements. *The first key* is to establish an individual product requirement for each subordinate team member that will define a difference between success and failure for the subordinate's effort. *The second key* is for performance management; each product selected must impact on the success or failure of the whole team effort.

The very structure of our modern military is seen to be marginal. There are entire military elements, such as information management or military supply, that are being manned without any essential product supporting a team performance. These elements become very inefficient, sucking up military dollars without assuring the support that other military elements need to perform at the highest level of effectiveness.

Internal support groups are not measurable performance units, as they have no product that must be provided for the larger performance to be a success. To be a part of the team, their products must meet the support needs of those who have key performance product responsibilities. The efficient support element will be manned to meet the needs of other units, and will be a success when those they support never have to pause in their central performance due to lack of support.

As the third key, the lead-officer will do all that he or she can to promote the success of each subordinate team member. Performance teaming starts with the leader who brings the team into existence; only when the performance leader starts his/her part of the performance effort first, can that leader reasonably call upon others to do their part.

The performance-leader's effort will include learning (for each subordinate) a performance process that will probably assure that success, and then teaching it to the one from whom performance will be required. It will include assuring that the subordinate will have the time, and other resources necessary to be a success. It will include personal interaction with the subordinate in a give-and-take environment to assure that the subordinate knows what is being required, and accepts that the resources provided (including the subordinate's own time and effort) will be sufficient to gain the result.

It is the action of the lead officer that establishes the team; without that first action, only a common team will form.

If some subordinate is already the accepted performance expert in an effort to be assigned, this interaction will have to start even earlier; the leader will have to learn the process from the team member. Only then can the leader truly say that he/she knows what leadership must provide to assure the success of that subordinate effort. Performance-leadership is active, not passive in its work efforts.

For the fourth key, the leader must instill performance attitude in the subordinate. This is a matter of trust. It is a trust that starts with the performance of the lead officer, who has already spent time and effort to plan the work; and who is spending time and effort in interaction with the subordinate to assure the productive result. It is a trust that the performance leader is going to be there to step in at need in order to assure that his/her people are successful in whatever is assigned.

It is a trust that other working officers on the team will receive this same level of support, and that each officer will accordingly be assured of others doing their part of the effort. It is a trust that the output from other officers on the team will be sufficient to meet the officer's input needs—both because it has been planned, and because the leader is there to actively make it happen in coordination with others. Even as in the soldier team, the officers on the leader's performance team are able to rely upon each other to do their part; this is even reliance upon the leader to do team management.

As the final key this becomes a performance team effort when the team members can rely upon each other because they all serve the same product-driven purpose. They are just meeting it through an organized effort that the leader has designed and initiated. The leader is not above the team; the leader is the heart of the team. A performance team will only come into being by the leader doing the work of team management, and being reliable in his/her willingness and ability to support the performance of other team members through further efforts.

Internal support efforts can also be arranged in this process. The base knowledge is that the performance members of the team will be able to assure the product through what they are doing.

The technique starts with a singular understanding of the team product. For a supporting element of the team, there is recognition that this product will be assured using an internal-support effort that provides all that other team members need to succeed in their functions. Success will not be defined by the support that the supporting member provides; it will be determined by meeting the support needs of others on the team.

With this, the same sense of reliance can be gained for internal support as for performance members. This team-approach establishes a definition of personal success for the support effort that is tied to the same performance product. There is a singular productive result that defines success for the larger effort, and each team member will be relying upon other members to succeed in their part of the effort. The lead officer will be there to handle variations; and other members of the team can rely upon their performance-leader to do this because that leader has the same product responsibility as the rest of the team.

28.5 Performance Team Maintenance

A stable team requires maintenance. There are two leadership efforts that are promised to those on the team. The first is general management. It is the continued planning of performance, along with a positive effort to maintain and improve performance processes and otherwise to assure the efforts of subordinate team members. It is a promise to interfere whenever there is some sort of challenge to the successful performance of any team member. It is a promise to seek out and adjust performance and support processes to best assure the performance output.

It is also a promise to work with individual team members whenever and however is necessary to best assure performance. It is an active personal interest in everything that promotes the successful performance of those on the team.

The second is exception management. This addresses an effort that arises to meet external nonperformance demands placed upon the team. It may be a change in performance requirement. It may be a direction to "demonstrate military performance" to visitors. It may be a death in the family of one of the team members.

Exception management is the promise of the leader to take positive and directive action to do, so far as possible, what will continue to assure the success of the working team in its singular productive purpose. If there is a change in performance requirement, the leader's self-assigned effort adjusts the planning to address this to others on the team in terms of their processes and individual performance requirements. If it is to deal with visitors, then the leader will take action to minimize any threat to team performance. If it is a death in a member's family, it is the team leader who takes action to gain permission for the team to take care of its own as a planned effort.

28.6 Performance from the Top

This concept of team starts with a lead-officer who determines to accept performance responsibility, and to enter into teaming with his/her subordinates. If this is a general officer, then it starts near the top of the larger organization.

With our example General, we assume a number of subordinate military organizations, each with its own local performance requirement. The first question is how the General is going to be able to initiate performance teaming at his level with these subordinates.

The effort starts with a clear understanding of a product requirement that the General accepts as a personal responsibility. Let us assume that it is the securing of a nation that is being assisted to political independence based on establishment of a peaceful democracy. The situation is that of a military overthrow of the previous government with the aid of the American military. There is chaos controlled by the military presence.

There will be internal steps or parts that "must" be achieved in fulfilling that responsibility—and these will define the subordinate products that are used to identify subordinate leader responsibilities. The General might recognize a first responsibility as the peace-keeping function which would include the establishment of a capacity in the local government to become independent in this function. The second would be public services, and would include training locals to take over this function. The third would be social and political services, generally the securing of the benefits of liberty for the people of the nation, both as an interim performance and a passing of these potentials to locals.

The final performance function would go to the leadership in the forming democracy. Yes, they would have to be part of the team if it was to succeed. This is why teaming is so important to the success of the effort.

The General sits down with his planners, and begins to delve into how each effort can succeed. In this, he brings in a subordinate to work with him (personal interaction) to develop a plan of action for the security and justice fields. This concentrates on what must happen, and what this subordinate will need to make it happen. It will include inputs in terms of American soldiers and equipment. It will include inputs in terms of local soldiers/police and equipment. The direction of the effort is to answer the general management question—If I can provide you with this, is there a good expectation that you can succeed at what I am asking of you?

There is a similar meeting with another subordinate concerning physical infrastructure. A third meeting addresses political and social liaison, and what the General should be asking from the Department of State as their part in the effort. Another meeting will be with a subordinate who will be the liaison to the new government for political needs.

When these plans are roughed out, then it is time to go to the Department of State, recognizing that there are political specialties and social services where the military should not be taking the lead. The general, by laying out the plans developed to that point, is able to address specific objectives and potential means for achieving these ends. The State Department representative, as the expert in this arena, can then work to accept, modify or replace the plans with a better plan where this DoS officer can give some reasonable assurance as to the result if the resources are provided. The DoS and liaison officer then function as a unit member of the performance team.

A similar meeting with leaders in the new government, addressing what the military and political leaders see as reasonably necessary to assure their part of the larger effort, address what they will need to take over on the general completing his direct involvement. Again, there is the initial plan to be considered, with the addition of expertise in the forming political government on what can reasonably be provided by the government in support of this effort.

The primary performance members for the team are then in place. The trust in one another due to the planning and assurances can be established if the resources are provided as promised.

They share the one productive purpose, and each is going to have to rely upon the others to make it happen.

A new officer is brought onto the team. This is the one who will provide the soldiers as support for the direction of various team members. This is the one who will transport and house soldiers and sailors, and provide them with the usual military command and support structure that keeps them healthy, on site, and ready for duty.

The officer who is there to address infrastructure develops the initial plan for action. If possible, a local government liaison is also brought on to work directly with this officer when it comes to planning. They become a two-person performance unit, and part of the team.

And the result of this effort is more than an excellent and workable plan. It is more than a living plan that is in charge of people who know when and how to modify it to meet changing needs and situations. What you have is the performance team, a structured working organization where each person has something to accomplish, has the basic resources to make it happen, and has reliance upon other members of the team for support. And in the center of this team, we find the General, who has been an active part of establishing all of the plans, and who has a high-level understanding of how each was derived, and the resourcing effort that provides the reasonable basis for success.

Better still, the members of this operating team know that they can rely upon that general because he has already demonstrated his commitment through being the initiator and through his active participation in the efforts that created the team. If they have exception needs, or new challenges arise, they know where to go to have them resolved. It is the General who is in charge of the team.

This is no common team. These are people who will work together because they have to work together to assure a success in the larger effort.

The technique uses a simple logic where each subordinate product requirement becomes a definition of success for some subordinate officer, specifically recognizing their relationship to the success of the General's larger effort. The success of the General is then openly tied to the success of these subordinate leaders.

It is then a short step to insisting that these subordinates also assure the performance in their areas through performance teaming. That is pretty much accomplished by the General's example; he has provided the performance-team model at his level, and assured that subordinates understand that this teaming process is how they can assure the performance of their subordinate areas. This should be seen as the leader teaching an efficient process to those who will be required to perform.

The most painful part of this effort will be handling internal support efforts within the senior-level team. Bringing these support managers into the attitude that they will be teaming with subordinate performance managers flies in the face of modern military thinking. The fact that others will rely upon them to provide the support that is needed rather than working with an output requirement for their support, is a new and challenging concept. Interactive planning with these support providers may be difficult; and it may require a team effort (including a gathering of performance members) to assure that there is a real sense of team. The General may expect to be an active and immediate leader in such a gathering, as subordinate performance officers may also have become convinced that a level-of-support effort will be adequate. Adequate internal-group output is not a team concept; only assured performance will establish the team.

In short, performance teaming at the higher administrative level can encourage performance teaming at subordinate levels. This can become part of the reliance that one subordinate officer puts upon another, that high-level team members will head a subordinate performance team that will be able to assure their local product.

29

How to Initiate Performance Management Within the US Army

Jesse W. Brogan
Walter Reed Army Institute of Research

29.1 Background

Management engineering is the application of basic industrial engineering in the office environment. It does for management and internal support what traditional industrial engineering has done for efficiency in productive efforts. Application to Army management provides a means to accomplish a larger Governmental purpose (performance management) in an area uniquely equipped to do the initial application.

This is no add-on management improvement program; it initiates a change that is effective at the level of organizational culture. It discovers an existing performance-management within the Army, and expands it from its present encapsulated area to address Army administration in general.

29.2 Challenge

The Federal Government mandate is performance-based management. The modern Army's management system is not performance-based, and appears actively hostile to performance basis. Attempts to bring performance-based management into areas controlled by Army management have yielded reduced performance while retaining inefficient and expensive administrative structures and processes. The challenge facing the Army is to keep its current effectiveness, while it institutes performance-based management.

29.3 Meeting the Challenge

This is a recommendation for correction, for instituting performance-based management where it is now missing. This is to be accomplished through basic management engineering, the application of efficiency engineering principles to the operation of an organization. This chapter identifies and defines the challenge in terms amenable to application of a cure.

1. It addresses the general direction necessary for accomplishing that cure.
2. It provides specific examples of applications that can increase effectiveness while reducing organizational costs.
3. It links this cure to current management practices in the Army that can be applied with assured effect.

Beyond this, the cure is general and instituted through potent understandings that are already well accepted for dealing with similar problems in management. These understandings can reflect back into the areas where the modern Army is already highly effective. There are additional performance-area benefits from the application of basic engineering to office-area operations.

29.4 Challenge Background

For foundation, the Army management system is loosely based on the work of Henri Fayol, who presented his management understandings in the early decades of the 20th century. His efforts were directed to administrative necessities for gaining performance, not to performance itself.

The challenge that faces the Army approach to management is fundamental, beginning with the definitions used for management and its associated work efforts. Army management is now defined by five functional activities—planning, organizing, coordinating, directing and controlling.

For performance perspective, it is easily possible to do all of these without producing anything. These are basic activities undertaken by managers in fulfillment of their duties, but are not performance determinates.

In the current Army approach, leadership is distinguished from management by the following: *Management is doing things right; leadership is doing the right things.* Both doing things right and doing the right things are also possible without gaining performance. This is especially true when "doing the right things" is defined to be performing the five processes, rather than gaining a performance through others.

The challenge can be restated: the Army management system has no foundation in performance. In general terms, Army managers above the first-line supervisory level are only rarely held accountable for gaining anything through the efforts of the organization over which they have direct authority. There is no effective performance-management.

This becomes clear when asking the functional question—What is the measurable difference between an Army management success and a management failure? The lack of any clear answer encapsulates the challenge that this paper addresses.

As a practical example, there have been major contracts for performances that have doubled in cost. Yet holding the military leader personally accountable to the extent of being taken out of authority over future contracted efforts is rare. The current system tolerates major performance failures!

Inability to gain performance is not a problem in itself. The failure to gain performance through the system of management controls and authorities is just a symptom of a deeper problem within Army management.

29.5 Performance Background

Performance-based management has its first foundation in the work of Frederick Taylor, a contemporary of Henri Fayol. He defined management as "gaining performance through the efforts of others" which is similar to the definition used by Fayol. Taylor's work was applied to production areas with great success.

Due to the nature of the times, Fayol's approach was much more attractive for meeting management requirements, especially in bureaucratic areas that had no obvious products that could define successful performance. Fayol's principles of administration were eventually applied to larger administrative efforts with good effect, yielding organizations that were able to control massive amounts of resources and keep their use in reasonable coordination.

Fayol's work was especially attractive to organizations like the military that had to have rigid authority structures to fulfill its primary mission. Fayol's work effectively addressed administration as performance of four administrative functions.

Administrative management was further defined by Fayol's 14 principles. Early administrative management did not consider itself to be subject to performance-management purposes, and did not apply Taylor's performance-management principles.

29.6 Corrective Activity Requirement

We are not addressing any self-healing situation. Existing senior management in the military has no effective orientation to organizational performance; and is not going to voluntarily promote management that is based on performance.

Correction will involve a proactive and designed effort to establish a new management foundation, one that is at least friendly to performance management. At a minimum, there should be someone who is personally accountable for each and every performance requirement placed upon any Army element; and the Army manager will have to be given control over the resources necessary to gain that performance. That same person will have to be held personally accountable for the success or failure of the performance.

29.7 A Viable Direction for a Cure

A managed performance arrangement is already the case when it comes to running military campaigns. Every military action is given to someone who accepts responsibility for a performance. In a military action, an officer is given a specific military objective, with defined results to be obtained through the efforts of those who answer to his or her command.

The challenge faced in this chapter can then be redefined. The challenge is expanding the Army understanding of "how to run an effective military campaign" into "how to operate an organization".

This direction of correction is both possible and is likely to yield the desired result.

There are two aspects to this new direction. The first is technical, addressing how it is done. The second is motivational, addressing how it can become something that military officers will implement.

The second appears impossible only when addressed in terms of modern Army management, which has no existing performance foundation. It becomes possible when performance management is required along with sufficient external stimulus to make it happen.

The first direction for the cure is to show how it can be done, including how internal and external Army leadership will address the cure to those who will implement it. The second is to provide the impetus that will get leadership to do what must be done.

29.8 Technical: How It can be Done

The keys are found in examining the necessary activities associated with managing a military campaign. Each commanding officer is given something that must be accomplished; and directive control over the resources that will be used for making it happen. Failure to accomplish the assigned results using the resources provided will threaten the accomplishment of the plans of the next-senior officer; and what that next-senior officer must accomplish will support the plans of the one in charge of the military campaign.

The first key is that sense of value that connects the activity of the lowest-level officer with the accomplishment of the larger efforts of which it is a part. In this environment, a failure to meet even a smaller objective can threaten failure for the larger organizational effort. Each officer is enlivened by the sense of importance, a feeling that the success of the larger military effort really can depend upon the performance of the local mission.

This defines one important part of our curative action, the discovery of "what must be accomplished" for the Army to be a success, expressed in terms that can be communicated down to the smallest part of the larger organization.

29.9 Motivational: How to Address this to Subordinateleaders

The potency of the general theory applied above is obvious when communicating performance-management to subordinates. It will be communicated by the same means that are used in communicating performance requirements in a military campaign. Once there is a definition of "what must be accomplished" by the Army, then the rest of the effort follows naturally. It includes developing and assuring the method of performance, and identification and dedication of the resources necessary to assure it.

This will institute a different motivational attitude than that now encountered in administrative management over the Army.

Specifically, if senior military leadership has discovered what must be done for the Army to be a success, subordinate officers don't get asked to buy-in to a senior-officer's military plans. Command directs performances that will get it done. While a leader may tap subordinate expertise for the best way to assure the performance, the result will still be directives that assure accomplishment.

29.10 Performance as a New Foundation

Management engineering is the application of basic efficiency engineering to the efforts that gain organizational performance; it provides key understandings that function as management tools. One such tool is the need for a new concept of what the Army must accomplish.

Management engineering is the direction for solution. It results from the application of Frederick Taylor's approaches, but applies these to the structuring and operation of organizations, with specific attention to efficiency of performances in the office environment.

For scoping, management engineering is engineering rather than management; it does not perform management. It provides managers the tools that will be most effective for gaining accomplishments through others. It is a support specialty.

As engineering, it is problem-oriented, but it is also technical-professional support for those managers who will make the decisions, and will take the actions that implement a solution to management problems.

29.11 The First Rule of Management Engineering

Management is an essential; You cannot improve management by replacing it with something else.

With management defined as "gaining a performance through others", this rule puts limits on activities that will be included in the efforts that will yield a systematic orientation to performance management.

As an initial efficiency concept:

There are only two purposes for any organizational element in an efficient organization. The first is performing mission. The second is supporting those who perform the mission.

Management is not part of performance; in an efficient organization, the efforts of managers will support performances accomplished by those who directly perform the mission.

Any effort that neither does organizational performance, nor supports that performance, can probably be terminated without loss. This even applies to administration. "Doing the right things" is not a separate specialty supporting activities by others. Accordingly, it is not to be approached a separate purpose for any leader. It is a part of administrative management, which exists only to fulfill the value-adding purpose of management.

Henri Fayol's process-based approach may be descriptive of what managers do, but is inadequate for assuring efficient performances. The key to efficient management is not that it performs necessary processes like controlling and directing, but that it supports and/or adds to performances that are managed.

To put this into our military-campaign perspective, the efforts of each level of Army leader adds value by assuring subordinate performances. As a practical example, the senior officer in a military campaign provides the junior officer with all that is necessary to assure accomplishment of an assigned military result.

When addressing military actions, the leader in charge is directly responsible for the care, maintenance and expenditure of assigned military resources. The commander starts with the clear purpose defined by the military action that is the basis for putting someone into command. This commander assures that his fighting people are supported by everyone else under his command, all the way from supply and information efforts to medical services.

Performance management for the Army as an organization begins with the establishment of "a performance that must be accomplished" if senior Army leadership is to be a success. This performance mandate becomes the driver for a military effort that has to be efficient and effective. This same level of performance purpose is needed for performance-based administrative efforts.

29.12 Tool: Customer-product Analysis

The engineered tool for establishing these key products is "customer-product analysis". This technique recognizes the ultimate importance of a working relationship between the organization and those who assure its continuance.

Step 1: Follow incoming resources back to find the decision-makers who determine to provide the Army with resources. These decision makers are defined to be "functional customers".

Step 2: Identify what these functional customers both receive and value as a basis for their resourcing decisions. All things that meet these two criteria are, by definition, "functional products".

Functional products are "what must be produced by the organization" if is to continue in existence and remain healthy. The must-do mission of the organization is delivering functional products to functional customers.

In the terms used above, the organization will be successful when it earns its resource base through delivering functional products to functional customers; and is likely to fail if it does not. Delivering functional products to functional customers is the performance driver that can unify internal efforts within all levels of the larger organization.

The Army is funded by Congressional action; the "functional customers" of the Army are members of Congress. There are also secondary customers who have influence with these primary decision-makers. There are also indirect customers who give feedback to these same primary customers. Still every focus comes back to the members of Congress whose decisions will determine the resources available to the Army to continue its existence and effectiveness.

Management engineering approach continues by addressing functional customers and functional products in terms of customer investment. Effective organizational activity will require intimate knowledge of the value that the customer will receive, and of the cost of gaining that value.

"Investment potential" is inherent in the product delivered, and includes such intelligence as will identify resourcing to assure the health and welfare of the Army organization.

Important to this recognition of product, the usual operation of the military, including maintenance of weapons systems and personnel, becomes an internal product, a product that is not delivered to decision makers to earn income, but is maintained as a subordinate part of assuring military value to those who will fund the organization. The investment potentials that the Army provides to members of Congress are the real source of value delivered; and delivery of this value is how the Army will "earn" its operating resources.

This product is to be an effective plan for action, with investment of resources indicated as sufficient to assure planned results. Performance management is undertaken to gain planned results through directing those provided resources.

Primary products probably include a presentation of an ability to militarily threaten other nations who might be adverse to our national purposes, and to minimize military threat by other nations to our interests.

With this, our modern Army has a peace-time military mission to accomplish, and it will need resources to direct to mission accomplishment. The difference between organizational success and failure will be inherent in delivering functional products where they have to go to be effective. Instead of a military objective, this provides a performance objective for accomplishment through military management.

29.13 Engineered Tool: The Concept of a Performance Module

A performance module is itself an engineered tool; it is the building block of every efficient organization. Understanding this tool starts with the example of the smallest operating unit in a military action. This is a single leader in direct charge of a number of soldiers.

Key points include the singular purpose given to that "local manager" in gaining a military objective. It includes the concept of a team, which both includes the officer as a key team member, and recognizes that officer as the one in charge. The team has a singular military objective, with vision maintained by the local leader.

Key efficiency considerations include superior officer abstention. Superiors do not give alternative or multiple objectives to operating subordinates. They do not step in and redirect members of the working team while it is engaged in its assigned military performance. They do not add secondary missions that might distract the group from its primary purpose.

Superior-officer impact includes support for the action team, both moral and physical. The superior officer effectively teams with junior officers to perform a management function that supports performance by the soldiers. The superior and subordinate managers form a superior team with an assigned objective. It forms a separate module with its own unique performance purpose.

Efficiency results from bringing these same considerations into the design of the larger Army organization.

Application begins at the most senior level. The objective is delivery of functional Army products to appropriate functional customers. Implementing an approved plan gives performance objectives that are supported by a clear difference between successful and failed performance. These are objectives that must be met for the health and welfare of the military organization, even as military-action objectives must be met for the effective waging of a military campaign.

In a different perspective, the single leader who is made responsible to implement a performance-management objective will need to be given sufficient resources to succeed in delivering the desired results as a functional product.

This single leader will not be able to personally direct the large resource base needed to meet any major performance objective. There must be a structure for action.

Efficient structure is created based on internal product requirements. In any larger effort, there will be internal products that must be gained through subordinate leadership if the larger performance objective is to be reached. These, like the objective itself, provide a basis for establishing performance-based subordinate elements. An officer, a military manager or senior leader, can be made responsible for providing that necessary internal product; can be resourced to make it happen; and then can be held personally accountable for gaining the desired result by applying the provided resources.

Performance is based on a deliverable result that separates success from failure. Every performance part of the larger organization will share in this concept of success.

29.14 Engineered Tool: Vertical Division of Management

Another key to understanding the efficiency aspect of modules is approaching management as work that must be performed by someone. A module is a basic performance unit; and each module created by use of this logic is manageable by its inputs and outputs. It will be a success or a failure by its ability to deliver its product through use of the resources given to it for that purpose.

This vertically divides management responsibility—an effort that has proven very difficult in the current Army management structure. Once responsibility for gaining a defined product through assigned resources is given to a subordinate, any further responsibility for directing those within the subordinate area also passes to that subordinate leader. The assigning senior leader only retains responsibility for exception management, and does not interfere unless unforeseen or otherwise exceptional needs arise.

The senior leader is only directly responsible to manage his local resources, those that are not dedicated to a subordinate performance effort. Thus the work of management is only done once, and done by the most appropriate officer.

Selection of subordinate modules is a matter for the art of management with a few general rules. The first is the application of productive purpose. Wherever there is a product where delivery is necessary for success of the next larger productive result, there is a point for establishing a performance module.

Management is always local. If a subordinate function is physically remote, then it will need a local leader; and the arrangement of a module is appropriate. If there are a number of remote areas involved in a single subordinate product, this could well indicate an efficiency-based need to consider reorganization. To be personally accountable for a performance, a module manager must be in direct and effective charge of those who are engaged in assuring the module product.

Also, the module manager answers only to the next superior module manager, the one who needs the subordinate-module product to succeed in providing his or her own superior-module product. Any third-party authorities who might interfere with this operational relationship are to be exceptional, and to be managed as such.

29.15 Engineered Tool: Structure of the Module

Part of what makes using modules efficient is the ability to hold each module manager personally responsible for generating the desired output product. That happens only when sufficient resources are dedicated to that module manager to assure the performance. Relying upon support from third-party authorities is an exception, not a benefit.

Like that lowest level military unit, a performance module is a complete performance unit. All of its necessary internal support is part of the module, and the local module commander is there to assure that this support meets the needs of the module's working members—those who assure that the module product is delivered as required.

This is one aspect of how that local module commander is able to assure that his support organization does in fact render the support needed. There is no intervening authority that might have any other purpose. The focus of the module is on the product that it must produce to be a success in its own operation.

As a general efficiency rule, the purpose of any internal support effort within a module is to provide support. Support groups *never* put work requirements upon any other part of the organization. This is especially true for superior-module support impacting work in subordinate modules, which are themselves managed by product and resources consumed. This is so different from current Army-management thinking that separate emphasis is necessary.

In another viewpoint, the internal support of the module is support for the module only. It owes nothing to support efforts in superior modules. For efficiency, a support group serves only the module in which it exists.

The leader in a subordinate module is part of the superior module, and is a representative for his or her subordinate module. The subordinate-module assignment of product responsibility is also the personal management responsibility of this subordinate-module manager.

Every module is inherently manageable by its cost and performance. Performance management is part of its very structure.

29.16 The Army Administration Mission

Now we can start putting this together into a cogent structure. We have established an initial definition of what must be accomplished. This product includes a delivery of information that connects Army funding with an ability to threaten nations whose purposes are adverse to those of decision-making Congressmen, and an ability to minimize the threat of hostile actions by foreign nations.

Now things start to come into focus in more familiar terms. The requirement for this will include the design, development, purchase and maintenance of weapons systems, and the preparation and maintenance of manpower to use those weapons. It will include the ability to transport these weapons and people to the point of application, and to militarily manage the actions that will bring them to bear for a purpose valued by those decision-makers.

This brings us to the major difference from today's management efforts. The type of planning that is required to accomplish a combined Army-action and Army-administration mission is based on the specific goals that are valued by decision-makers. These decision-makers are representatives, and representatives may have to sell their decisions to their constituents. This is further guidance for how the Army must communicate value to these decision-makers. The Army plan for success will contain the communication of military value to the public, so that it can be presented by decision-making congressmen.

There are questions to be answered. What is necessary and convenient for a military establishment that will be able to contain and respond to foreign hostility? What level of flexibility should be planned into the Army, so that it can respond to unforeseen needs? What is an appropriate maintenance level for men and equipment? What will be required when the Army has to go from a maintenance state to one of military activity, and how much of the change will be funded up front to allow action before there is separate/exceptional Congressional support?

These are all questions to be answered by a performance plan; and that performance plan will be approached as a primary product for delivery to decision makers so that they fund the Army's continuation and effectiveness. That plan must be prepared at the level that will be deliverable to decision-makers, and will meet their needs for communication to their constituents. That plan will be a primary investment tools for the decision makers, through which they decide to assure proper resourcing for the Army.

There will be no separate plan for "security issues" unless these are to be separately funded security efforts. There may be a secured area within the general plan that will only be disclosed to decision-makers. Communication support for constituents may have to be adjusted accordingly. Such functional-product development will require the highest form of strategic management thinking.

The value of this approach to planning is obvious. The Army will have something that it "must accomplish" if it is to be a success; and it has the additional purpose of communicating this "must fund" concept to decision makers. Once money is committed, it is unlikely to ever be challenged or withdrawn so long as planned performance is achieved.

29.17 Distribution of Performance Responsibilities

When this new type of plan is resourced it provides a foundation for actions within the Army that "must be" accomplished if the Army is to be a success. This is development of the key component for performance-based Army management. Implementation of the plan up to the level funded contains an effective promise that the military leadership has made to its congressional decision-makers.

To make this general approach effective, the larger elements of the plan must be fulfilled by subordinate military authorities. There is likely to be a schedule of military hardware to be developed, purchased and maintained to meet potential performance requirements for that hardware. For performance management, some specific subordinate can be made personally responsible to accomplish this performance element; and be given control over planned resources with the purpose of assuring the result.

With this, we have a performance product that "must be delivered" if the Army is to expect future funding for its continued equipment-based efforts. This meets our requirement for performance management. The success of the superior effort is dependent upon the success of the subordinate.

The Army will have to maintain the readiness of officers and soldiers at a planned level, and this will be a measurable performance that can be assigned to a specific officer with responsibility to make it happen with resources provided. Again, the fulfillment of this responsibility is necessary if the Army is to be a success in the viewpoint of those whose decisions fund it.

29.18 The Depth of Change

This describes no minor change that can be gently accomplished. Changing from authority-based to performance-based management affects the culture of the Army. Where an authority-based system treats cost overrun in the purchase of a weapons system as a normal variation, the same would, in a performance-based system, be a performance failure for the one responsible to meet the performance objective. The planning (a senior management function) would have failed, and so would the planners. Someone would be accountable for the failure in performance; someone would be accountable for the failure in planning.

The cozy relationship between military planners and those in the military industrial complex would become much more business-oriented. Cost failure would become Army failure, witnessing a deeper failure in Army management.

As with a military campaign, the necessity of success at all levels would have to become a part of the operating environment. This was the goal to be accomplished in this paper. Part of the effect is assuring that senior leadership knows how to manage, to gain performance through direction of Army resources.

The logic is a continuation of the common sense of every commander. If any subordinate officer does not know how to be a success at gaining a performance result, they should not be given authority to act as subordinate leaders assigned to gain it.

This change is so basic that it will affect even the Army definition of leadership. Leadership will no longer be definable as the nonperformance "doing the right things". Leadership will be redefined to "gaining the desired outcome" through the efforts of the organization.

The separation of leadership from management—which is traditional for Army management—will no longer be tolerated. Officers who are unwilling to perform to the plan that is promised to Congress are likely to find themselves out of command. It will be an Army where the thinkers support the doers, instead of the other way around; and that is a major cultural change.

Switching the Army from an authority-based system that is tolerant of failures (such as cost overruns and manpower training weaknesses) is not going to be an easy shift; it is going to require substantial management leadership. Someone is going to have to plan that change, and present an intelligent investment plan to Congress for resourcing the changeover effort.

Reversing a segment of the military culture, even with the obvious example of military actions being within that same culture, is not going to be done without substantial consequences. These will include such potentials as abandonment of Army careers by those who find performance-based management unacceptable.

29.19 Change Management

The Army is uniquely able to handle this type of strategic planning. Army leadership has working parallels in the handling of casualties in military actions. Army leadership is specially prepared for the type of strategic thinking that yields a desired result in an area that involves great conflict and individual hardship. Their working tools for this are probably superior to those available from other sources.

The changeover begins by requiring a plan for the changeover. Once the changeover becomes the goal, the Army has the ability to find the commander to be in charge, and to begin the process of establishing the battle plan to make it happen.

Congress has a part to play in this action. The officer selected for this duty will have to be put into effective command over the action, and this is unlikely to occur without separate legislation. There are too many conflicting authorities in the existing Army management system. Other senior officers are likely to resist, or even to defeat, the change in continuation of present priorities. This officer has to be in authority to enforce discipline over all Army officers for the purpose of assuring the change.

Management engineering has an applicable general rule: *The one who is put in charge of the change is also to be put in charge of the resulting situation.* This officer will have to accept personal responsibility and accountability for running the organization after it is changed. This puts performance basis even upon the change. The one with authority to establish the change will end up being personally accountable for the result. Failure will not be a viable option.

29.20 Elements of Culture Change

One of the first casualties of the change will be management by regulation. In a performance environment, regulatory management will be approached as an aggravated form of micro-management. The subordinate area is managed by product and cost, not by process requirement. Again, the change will be made at a culture level.

This change is endemic, and the benefits can be presented by example through a short study of property management. The property management system for Fort Meade effectively tracked property in accord with regulation. It yielded once-a-year inventories, and required four full-time employees, and the part time equivalent of one full time employee in the effort performed by hand-receipt holder actions.

The performance-based alternative recognizes that the logistic purpose is support, not management control; and support work is *never* to be placed upon those who receive the support. The only reason to have a separated support group is for it to provide the support that others need.

The indicated property support effort requires action by the logistic experts. They go out and physically observe and track property. Signature tracking is still required, but the logistic experts are the ones who fill out their own internal logistic paperwork.

Two people could easily observe each piece of equipment six to ten times a year, and prepare all the logistic paperwork. This would multiply the ability to control and maintain Army property. It accomplishes this major improvement using only a fraction of the cost now expended on this one minor logistic-support function.

The more important effect is not in logistics, it is the lifting of logistic support work from those who are focused on organizational performance. Performance people can be more focused on generating the products that will make the larger organization a success. Performance management initiates a very different culture than that now in place.

It puts managers more in charge of performance, and increases their ability to assure performance through those working under their command.

This example is not unique, but is indicative of almost all internal support efforts that are controlled through regulation. The cost of regulatory management is inefficiency for both the support function and for the performance function that it exists to support.

The effect is also endemic. Consider the operation of an Army Base, now a separately management performance area managed from the level of an Assistant Secretary of the Army. It is indicative that there is no such thing as a measurable failure or success at operating army bases!

Management starts with performance; and Army bases in the United States do not provide a defined and measurable performance that must be accomplished if base management is to be a success.

In terms of our functional product, the domestic Army base does not threaten foreign nations with military action, and it does not interfere with their ability to threaten us. It is internal support. Base services are to be managed at the level of the performances that receive base-operations support.

The performance will belong to others. The base is a means for providing necessary and convenient internal support for meeting productive purposes. Bases may or may not be the most effective way of supporting performance.

Internal support efforts are subordinated to the performance areas they support. Support efforts within any module are to be missioned and funded by the leader of the module. Support resources should be determined by what the module's subordinate performance efforts need in order to assure the generation and delivery of their necessary performance products.

With performance-based management, attempted control by those higher in the organization would be micro-management. Army-level value is not delivered by managing a support organization, but by gaining products that are necessary if Army management is to be a success in delivering its functional products to its functional customers.

Even if an Army base is mandated from above as an Army-level investment, the base-level officers who have responsibility to operate these support organizations will have to answer to module managers in charge of the areas that receive support. This is necessary if performance managers are to be fully accountable for success in performing their military purposes; and support managers are to be fully accountable for their supporting efforts.

There is still a valid place for general regulation, but not for establishing centralized and mandated processes. The commander with a performance responsibility is also responsible for establishing and maintaining such internal support efforts as will meet the requirements of those subordinate performance areas under his immediate control. Regulation can provide intelligent guidance.

Also, centralized support efforts like military supply, can provide internal support at the centralized level. It is reasonable to offer support to subordinate commanders, and these subordinate commanders can offer the same to their subordinates.

If the support is truly supportive, it is likely that the subordinate commanders will welcome it. If not, other options should be sought out and implemented. The value of support for a local process is determined by those who receive the support; and this need is not generally known to managers above the local-module level. Their senior-level investment must be at their level, and should not interfere with the operation of subordinate modules that they have been entrusted to subordinate officers.

29.21 The Engineered Estimating Process

A final engineered tool is necessary for intelligent process, that for establishing the resource-cost information that will support the performance objective. This is the engineered tool for change management.

The management-engineering tool is a performance management concept for estimating. It starts with the performance concept and it follows its general logic and approach toward establishing a reasonable promise to perform.

The same general logic will be employed as that used by a military commander faced with a military objective. He has to energize and distribute his subordinate military resources to accomplish that central objective.

The military application starts with a clear statement of the objective, which will only be sufficient when it is clear enough to distinguish between a successful and failed military operation. Then the personal effort of that commander will succeed or fail as the subordinate operation under his command succeeds or fails.

The performance-based change effort for the Army starts with defining the functional product. That product will be valued by the decision-makers who will receive it. When there is a clear and measurable distinction between a successful and failed effort, then product delivery will adequately define successful performance; and successful performance will define success for performance management. Accepting the Army-level product as a performance responsibility establishes the largest performance module, the Army as a whole.

Visualization and planning that identifies, designs and delivers the value to decision-makers is the primary productive effort for Army management. That effort can either be assumed by the most senior manager, or be assigned to a subordinate for performance.

Ability to succeed at the military mission proceeds with an ability to determine subordinate mission objectives that will, if accomplished, lead to a successful military conclusion at the senior level. These are assigned to subordinate commanders. Subordinate commanders, following the same basic process, assign responsibility and accountability to their subordinates. The second-tier officer will address those subordinate objectives that will result in their successful performance of what the superior commander set them to do.

Each military commander is given charge of resources felt to be sufficient to accomplish the military objectives that are assigned. The most-senior commander endeavors to distribute available resources to subordinate commanders in support of their performances, reserving such as seem reasonable to handle exception situations that arise. Second-tier officers, of course, do the same for their subordinates, attempting to reasonably assure that their subordinates can accomplish what is directed so that the larger purpose is met.

This general logic is applied until all the military mission is planned for immediate action. While this is a continuing process based on command priorities, operating environment, interim results and resource capacities, there is a general and systematic logic to the application of resources to accomplish the military objective.

The Army, as an administrative organization, has a different challenge when it comes to resources. The resources are not fixed, even though they tend to be fairly stable. They are now distributed to subordinate areas in accord with a modifiable proposal generated for that purpose and blessed by Congress.

Performance management provides a different foundation, one which opens the potential for more effective planning and resourcing.

Estimation is the effect of having second-tier officers come back to the most senior manager with estimates of what it will cost to satisfy the performance objective that has been assigned. They, of course, have the same relationship with their subordinate performance managers. This process goes down to the most basic working element that has a performance responsibility.

The roll-up at each level adds the cost of the local resources needed for local management and administration to assure proper and effective support for all the assigned subordinate performance areas. To this, some reasonable amount is added in support of a level of expected changes and alterations that are likely to occur in the local operation.

Once performance responsibilities are set by this engineered estimating method, the establishment of resource requirements can be a fairly short process, a matter of a few days for each layer of performance responsibility. Then the performance will be assured by a known resourcing that is effective throughout the organization.

The only valid reason for returning to the decision-makers for increases after they have resourced the effort will be an exceptional amount of change, such as might be encountered if the Army is called into a major and active conflict.

29.22 The Final Word

This is not just a challenge for the Army administrative organization. The focus of administrative management upon performance has impacts that echo back to the operational side of the military. A clear functional knowledge of the flow of value can have great impact on the efficiency of military operations.

For example, the current military supply system is justly touted for its remarkable ability to deliver the goods to where they are needed. With the current military structure where supply is managed from a central location, the costs of achieving that end are remarkably well hidden. Decisions are not being intelligently made on the basis of cost and value.

When cost is examined, we find that much of the cost has been hidden in customer organizations, where every customer group has someone trained to perform supply transactions.

The time to train these people is a cost of supply administration. The time that these workers spend on gaining supplies and materials from the centralized supply system is hidden supply cost.

In a similar sense, there are unique financial and property tracking requirements that are placed on customer organizations. This is also a logistic cost, with effort incurred by customers in order to gain supplies. It is hidden cost of the supply system.

While it is difficult to quantify these costs, it appears that they are in excess of 5% of the manpower cost of the military. One out of every 20 people is a very stiff price to pay just to have the materials for war available when and where they are needed. For peace-time, it is unacceptable.

The challenge is in administrative management; and so the solution will also be in administrative management. Application begins with customer-product analysis. The purpose for the existence of this support organization is to meet the supply needs of those who have war-fighting responsibilities. The organization needs to get materials to the users when and where they need them to perform their war-fighting preparation and performance.

The general support rule is applicable; support organizations *never* put work requirements on working elements of the organization. They provide the support they exist to provide. We want the war-fighting elements concentrating their resources on fighting wars, not on getting their own supply. It is up to a supply-support group to maintain knowledge of support needs in serviced groups, and to satisfy them.

Management engineering provides a test for the success or failure of an internal support effort. A supply effort will be a success if there is no threat of failure in any operating element due to lack of adequate supplies and materials.

By this performance-based shift in administrative vision, resources for support are shifted out of the operating organizations, and into the support group. These groups are remissioned to provide the support that assures success in operating groups, rather than providing a way for these groups to get their supplies and materials.

The cost of providing supply support is likely to drop dramatically, even as the support is made more certain because accomplishing support becomes a logistic expert's measure of success. Beyond this, the working elements of the organization can be more effectively focused on what the larger organization must accomplish to be a success in its operation.

The larger effect of the shift of administrative management into performance mode is that it can also focus its own energies on supporting performance, rather than just trying to perform the functions of administration as its operation. The purpose of administration is subordinated to, and becomes part of, the performance purpose of the organization.

Management engineering puts the manager more in charge than is possible in authority-based systems, and assures that management adds value to what performance areas accomplish.

30

Critical Resource Diagramming and Work-Rate Analysis

Adedeji B. Badiru
Air Force Institute of Technology

30.1 Introduction

Real world projects are usually executed using limited resources. Therefore, decisions must be made about resource allocation in order to achieve project goals within the constraints of time, quality, and cost. During the past decades, several researchers [1,5–10,13,14] have developed heuristic and optimization methods for project activity scheduling, which take resource constraints into consideration. In addition, Badiru [2] proposed a graphical tool called critical resource diagram (CRD) to facilitate resource allocation in projects scheduling. A CRD is an extension of critical path method (CPM) and it has several advantages which include its simplification in resource tracking and control, better job distribution, improved information to avoid resource conflicts, and better resource leveling [3]. In this chapter, we formulate a mixed-integer program that is based on the concepts of CRD for solving resource scheduling problems.

Basic CPM and program evaluation and review technique (PERT) approaches assume unlimited resource availability in project network analysis. In realistic projects, both the time and resource requirements of activities should be considered in developing network schedules. Projects are subject to three major constraints: time limitations, resource constraints, and performance requirements. Since these constraints are difficult to satisfy simultaneously, trade-offs must be made. The smaller the resource base, the longer the project schedule. Resource allocation facilitates the transition of a project from one state to another state. Given that the progress of a project is in an initial state defined as S_i and a future state is defined as S_f, then three possible changes can occur.

1. Further progress may be achieved in moving from the initial state to the future state (i.e. $S_f > S_i$).
2. Progress may be stagnant between the initial state and the future state (i.e. $S_f = S_i$).
3. Progress may regress from the initial state to the future state (i.e. $S_f < S_i$).

Resource allocation strategies must be developed to determine which is the next desired state of the project, when the next state is expected to be reached, and how to move toward that next state. Resource availability and criticality will determine how activities should be assigned to resources to facilitate progress of a project from one state to another. Graphical tools can provide guidance for resource allocation strategies. CPM, PERT, and precedence diagramming method (PDM) are examples of graphical tools based on activity scheduling. Unfortunately, similar tools are not available for resource scheduling. There is a need for simple tools for resource allocation planning, scheduling, tracking, and control.

Figure 30.1 shows three resource loading options. In the first case, activity-resource assignments are made on a one-to-one basis. In the second case, one resource is assigned to multiple activities. There are also cases where a single activity may be assigned to more than one resource unit. The specific strategy used will depend on the prevailing resource constraints.

Resources are needed by activities, activities produce products, products constitute projects, and projects make up organizations. Thus, resource management can be viewed as a basic component of the management of any organization. It is logical to expect different resource types to exhibit different levels of criticality in a resource allocation problem. For example, some resources may be very expensive. Some resources may possess special skills. Some may have very limited supply. The relative importance of different resource types should be considered when carrying out resource allocation in activity scheduling. The CRD helps in representing resource criticality.

30.2 Resource Profiling

Resource profiling involves the development of graphical representations to convey information about resource availability, utilization, and assignment. Resource loading and resource leveling graphs are two

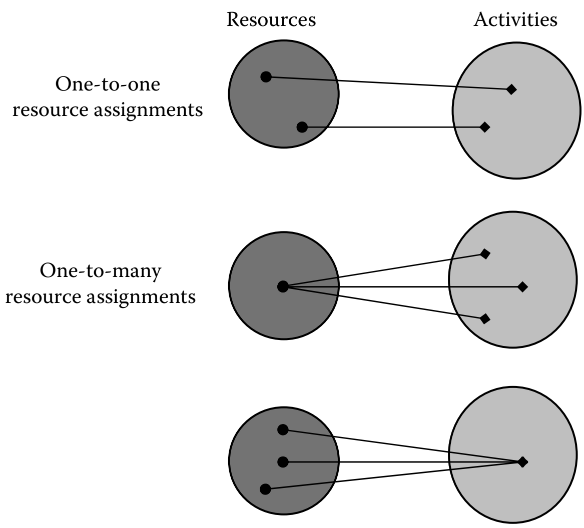


FIGURE 30.1 Activity-resource assignment options.

popular tools for profiling resources. Resource idleness graph and CRD are two additional tools that can effectively convey resource information.

30.3 Resource Loading

Resource loading refers to the allocation of resources to work elements in a project network. A resource loading graph presents a graphical representation of resource allocation over time. Figure 30.2 shows an example of a resource loading graph. A resource loading graph may be drawn for the different resources types involved in a project.

The graph provides information useful for resource planning and budgeting purposes. In addition to resource units committed to activities, the graph may also be drawn for other tangible and intangible resources of an organization. For example, a variation of the graph may be used to present information about the depletion rate of the budget available for a project. If drawn for multiple resources, it can help identify potential areas of resource conflicts. For situations where a single resource unit is assigned to multiple tasks, a variation of the resource loading graph can be developed to show the level of load (responsibilities) assigned to the resource over time. Table 30.1 shows an example of a resource availability database for drawing a resource loading graph.

30.4 Resource Leveling

Resource leveling refers to the process of reducing the period-to-period fluctuation in a resource loading graph. If resource fluctuations are beyond acceptable limits, actions are taken to move activities or resources around in order to level out the resource loading graph. Proper resource planning will facilitate a reasonably stable level of the work force. Advantages of resource leveling include simplified resource tracking and control, lower cost or resource management, and improved opportunity for learning. Acceptable resource leveling is typically achieved at the expense of longer project duration or higher project cost. Figure 30.3 shows a somewhat leveled resource loading.

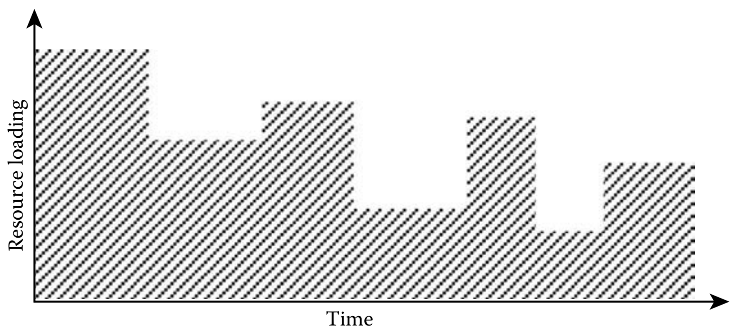


FIGURE 30.2 Initial resource loading graph.

TABLE 30.1 Example of Resource Availability Database

Resource ID	Brief Description	Special Skills	When Available	Duration of Availability	How Many
Type 1	Technician	Electronics	8/05/2009	2 months	15
Type 2	Programmer	Database	12/01/2009	Indefinite	2
Type 3	Engineer	Design	Immediate	5 years	27
⋮	⋮	⋮	⋮	⋮	⋮
Type *n*-1	Operators	Data Entry	Always	Indefinite	10
Type *n*	Accountant	Contract laws	09/02/2009	6 months	1

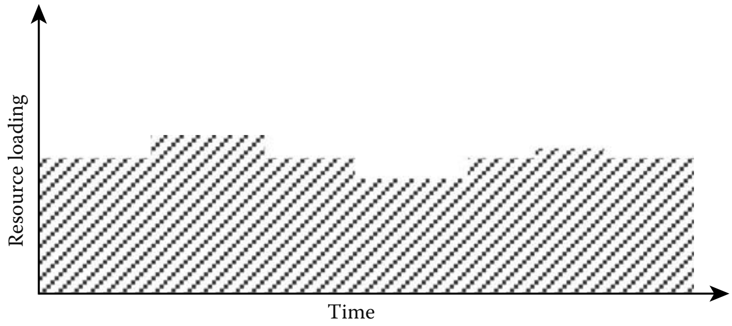


FIGURE 30.3 Leveled resource loading.

It should be noted that not all of the resource fluctuations in a loading graph can be eliminated. Resource leveling attempts to minimize fluctuations in resource loading by shifting activities within their available slacks. One heuristic procedure for leveling resources, known as the Burgess's method [2], is based on the technique of minimizing the sum of the squares of the resource requirements in each period over the duration of the project.

30.5 Resource Idleness

A resource idleness graph is similar to a resource loading graph except that it is drawn for the number of unallocated resource units over time. The area covered by the resource idleness graph may be used as a measure of the effectiveness of the scheduling strategy employed for a project. Suppose two scheduling strategies yield the same project duration and a measure of the resource utilization under each strategy is desired as a means to compare the strategies. Figure 30.4 shows two hypothetical resource idleness graphs for the alternate strategies. The areas are computed as follows:

$$\text{Area A} = 6(5) + 10(5) + 7(8) + 15(6) + 5(16)$$

$$= 306 \text{ resource-units-time.}$$

$$\text{Area B} = (6) + 10(9) + 3(5) + 6(5) + 3(3) = 12(12)$$

$$= 318 \text{ resource-units-time.}$$

As can be seen from the preceding sections, resource management is a complex task that is subject to several limiting factors, including the followting:

- Resource interdependencies
- Conflicting resource priorities
- Mutual exclusivity of resources
- Limitations on resource substitutions
- Variable levels of resource availability
- Limitations on partial resource allocation
- Limitations on duration of availability

Several tools have been presented in the literature for managing resource utilization.

Badiru [2,3] introduces the CRD as a tool for resource management. Figure 30.5 shows an example of a CRD for a project that requires six different resource types. The example illustrates how CRD can be used to develop strategies for assigning activities to resources and vice versa. Each node identification, RES *j*, refers to a task responsibility for resource type *j*. In a CRD, a node is used to

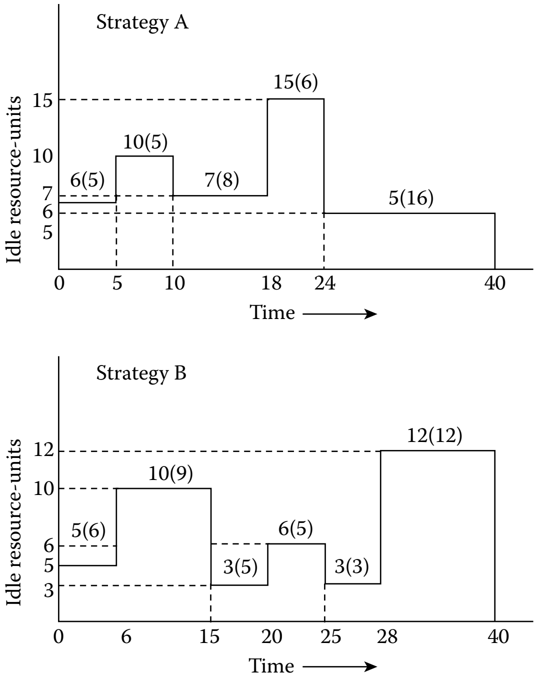


FIGURE 30.4 Resource idleness graphs for resource allocation strategies.

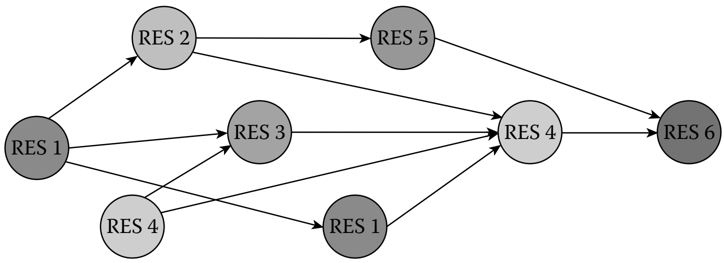


FIGURE 30.5 Basic critical resource diagram.

represent each resource unit. The interrelationships between resource units are indicated by arrows. The arrows are referred to as *resource-relationship (R-R) arrows*. For example, if the job of resource 1 must precede the job of resource 2, then an arrow is drawn from the node for resource 1 to the node for resource 2.

Task durations are included in a CRD to provide further details about resource relationships. Unlike activity diagrams, a resource unit may appear at more than one location in a CRD provided that there are no time or task conflicts. Such multiple locations indicate the number of different jobs for which the resource is responsible. This information may be useful for task distribution and resource leveling purposes. In Figure 30.6, resource type 1 (RES 1) and resource type 4 (RES 4) appear at two different nodes, indicating that each is responsible for two different jobs within the same work scenario.

However, appropriate precedence constraints may be attached to the nodes associated with the same resource unit if the resource cannot perform more than one task at the same time. This is illustrated in Figure 30.6.

In an application context, CRD can be used to evaluate the utilization of tools, operators, and machines in a manufacturing system. Effective allocation of these resources will improve their utilization levels. If

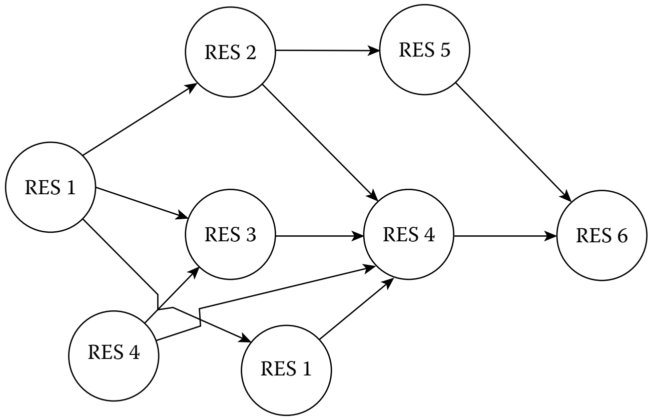


FIGURE 30.6 CRD with singular resource precedence constraint.

tools that are required at several work sites are not properly managed, bottleneck problems may develop. Operators may then have to sit idle waiting for tools to become available, or an expensive tool may have to sit unused while waiting for an operator. If work cannot begin until all required tools and operators are available, then other tools and operators that are ready to work may be rendered idle while waiting for the bottleneck resources. A CRD analysis can help identify when and where resource interdependencies occur so that appropriate reallocation actions may be taken. When there are several operators, any one operator that performs his/her job late will hold up everyone else.

30.5.1 CRD Network Analysis

The same forward and backward computations used in CPM are applicable to a CRD diagram. However, the interpretation of the critical path may be different since a single resource may appear at multiple nodes. Figure 30.7 presents a computational analysis of the CRD network in Figure 30.5. Task durations (days) are given below the resource identifications. Earliest and latest times are computed and appended to each resource node in the same manner as in CPM analysis. RES 1, RES 2, RES 5, and RES 6 form the critical resource path. These resources have no slack times with respect to the completion of the given project. Note that only one of the two tasks of RES 1 is on the critical resource path.

Thus, RES 1 has a slack time for performing one job, while it has no slack time for performing the other. None of the two tasks of RES 4 is on the critical resource path. For RES 3, the task duration is specified as zero. Despite this favorable task duration, RES 3 may turn out to be a bottleneck resource. RES 3 may be a senior manager whose task is that of signing a work order. But if he or she is not available to sign at the appropriate time, then the tasks of several other resources may be adversely affected. A major benefit of a CRD is that both the senior-level and lower-level resources can be included in the resource planning network.

A *bottleneck* resource node is defined as a node at which two or more arrows merge. In Figure 30.7, RES 3, RES 4, and RES 6 have bottleneck resource nodes. The tasks to which bottleneck resources are assigned should be expedited in order to avoid delaying dependent resources. A *dependent* resource node is a node whose job depends on the job of immediate preceding nodes. A *critically dependent* resource node is defined as a node on the critical resource path at which several arrows merge. In Figure 30.7, RES 6 is both a critically dependent resource node and a bottleneck resource node. As a scheduling heuristic, it is recommended that activities that require bottleneck resources be scheduled as early as possible. A *burst* resource node is defined as a resource node from which two or more arrows emanate.

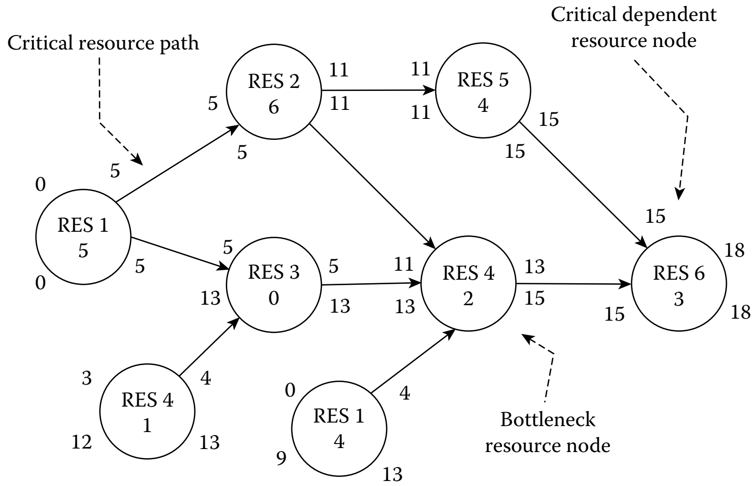


FIGURE 30.7 CRD network analysis.

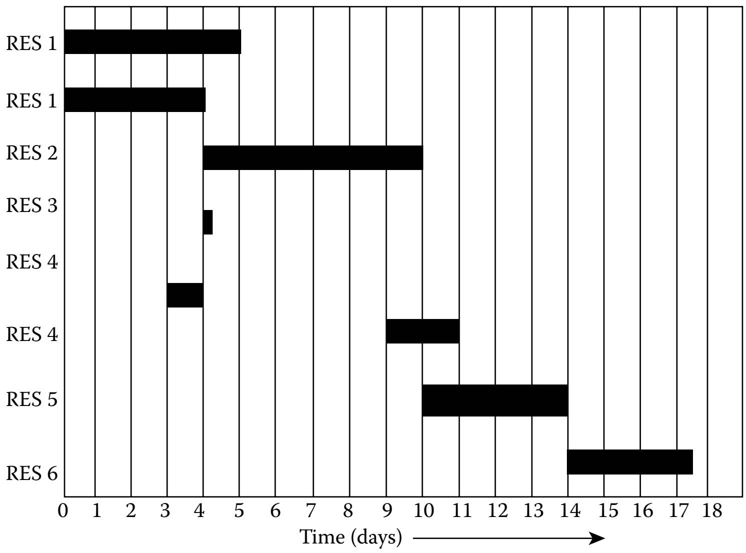


FIGURE 30.8 Resource schedule chart based on earliest start times.

Like bottleneck resource nodes, burst resource nodes should be expedited since their delay will affect several following resource nodes.

30.6 Resource Schedule Chart

The CRD has the advantage that it can be used to model partial assignment of resource units across multiple tasks in single or multiple projects. A companion chart for this purpose is the resource schedule (RS) chart. Figure 30.8 shows and example of an RS chart based on the earliest times computed in Figure 30.7. A horizontal bar is drawn for each resource unit or resource type. The starting point and the length of each resource bar indicate the interval of work for the resource. Note that the two jobs of RES 1 overlap over a four-day time period. By comparison, the two jobs of RES 4 are separated by a period of six days. If RES 4 is not to be idle over those six days, "fill-in" tasks must be assigned to it. For resource jobs that overlap, care must be taken to ensure that the resources do not need the same tools (e.g. equipment, computers, lathe, etc.) at the same time. If a resource unit is found to have several jobs overlapping over an extensive period of time, then a task reassignment may be necessary to offer some relief for the resource. The RS chart is useful for a graphical representation of the utilization of resources. Although similar information can be obtained from a conventional resource loading graph, the RS chart gives a clearer picture of where and when resource commitments overlap. It also shows areas where multiple resources are working concurrently.

30.7 Resource Work Rate Analysis

When resources work concurrently at different work rates, the amount of work accomplished by each may be computed by a procedure presented in [4]. The CRD and the resource schedule chart provide information to identify when, where, and which resources work concurrently. The general relationship between work, work rate, and time can be expressed as

$$w = rt$$

where w = amount of actual work accomplished (expressed in appropriate units, such as miles of road completed, lines of computer code typed, gallons of oil spill cleaned, units of widgets produced, and surface area painted. r = rate at which the work is accomplished. t = total time required to accomplish the work.

It should be noted that work rate can change due to the effects of learning curves. In the discussions that follow, it is assumed that work rates remain constant for at least the duration of the work being analyzed.

Work is defined as a physical measure of accomplishment with uniform destiny (i.e. homogeneous). For example, a computer programming task may be said to be homogeneous if one line of computer code is as complex and desirable as any other line of code in the program. Similarly, cleaning one gallon of oil spill is as good as cleaning any other gallon of oil spill within the same work environment. The production of one unit of a product is identical to the production of any other unit of the product. If uniform work density cannot be assumed for the particular work being analyzed, then the relationship presented above will need to be modified. If the total work to be accomplished is defined as one whole unit, then the tabulated relationship in Table 30.2 will be applicable for the case of a single resource performing the work, where $1/x$ is the amount of work accomplished per unit time. For a single resource to perform the whole unit of work, we must have the following:

$$\frac{1}{x}(t) = 1.0$$

That means that magnitude of x must equal the magnitude of t. For example, if Machine A is to complete one work unit in 30 min, it must work at the rate of 1/30 of work per unit time. If the magnitude of x is greater than the magnitude of t, then only a fraction of the required work will be performed. The information about the proportion of work completed may be useful for resource planning and productivity measurement purposes. In the case of multiple resources performing the work simultaneously, the work relationship is as shown in Table 30.3.

TABLE 30.2 Single Resource Work Rate Data

Resource	Work Rate	Time	Work Done
Machine A	$1/x$	t	1.0

TABLE 30.3 Multiple Resource Work Rate Data

Resource Type i	Work Rate, r_i	Time, t_i	Work Done, w_i
RES 1	r_1	t_1	$(r_1)(t_1)$
RES 2	r_2	t_2	$(r_2)(t_2)$
…	…	…	…
RES n	r_n	t_n	$(r_n)(t_n)$
		Total	1.0

For multiple resources, we have the following expression:

$$\sum_{i=1}^{n} r_i t_i = 1.0$$

where n = number of different resource types; r_i = work rate of resource type i; t_i = work time of resource type i.

The expression indicates that even though the multiple resources may work at different rates, the sum of the total work they accomplished together must equal the required whole unit. For partial completion of work, the expression becomes:

$$\sum_{i=1}^{n} r_i t_i = p$$

where p is the proportion of the required work actually completed. Suppose that RES 1, working alone, can complete a job in 50 min. After RES 1 has been working on the job for 10 min, RES 2 was assigned to help RES 1 in completing the job. Both resources working together finished the remaining work in 15 min. It is desired to determine the work rate of RES 2.

The amount of work to be done is 1.0 whole unit. The work rate of RES 1 is 1/50 of work per unit time. Therefore, the amount of work completed by RES 1 in the 10 min it worked alone is (1/50)(10) = 1/5 of the required work. This may also be expressed in terms of percent completion or earned value using C/SCSC (cost-schedule control systems criteria). The remaining work to be done is 4/5 of the total work. The two resources working together for 15 min yield the results summarized in Table 30.4.

Thus, we have $15/50 + 15(R_2) = 4/5$, which yields $r_2 = 1/30$ for the work rate of RES 2. This means that RES 2, working alone, could perform the job in 30 min. In this example, it is assumed that both resources produce identical quality of work. If quality levels are not identical for multiple resources, then the work rates may be adjusted to account for the different quality levels or a quality factor may be introduced into the analysis. The relative costs of the different resource types needed to perform the required work may be incorporated into the analysis as shown in Table 30.5.

As another example, suppose that the work rate of RES 1 is such that it can perform a certain task in 30 days. It is desired to add RES 2 to the task so that the completion time of the task could be reduced. The work rate of RES 2 is such that it can perform the same task alone in 22 days. If RES 1 has already worked 12 days on the task before RES 2 comes in, find the completion time of the task. It is assumed that RES 1 starts the task at time 0. As usual, the amount of work to be dome is 1.0 whole unit (i.e. the full task). The work rate of RES 1 is 1/30 of the task per unit time and the work rate of RES 2 is 1/22 of the task per unit time. The amount of work completed by RES 1 in the 12 days it worked alone is (1/30)(12) = 2/5 (or 40%) of the required work. Therefore, the remaining work to be done is 3/5 (or 60%) of the full task. Let T be the time for which both resources work together. The two resources working together to complete the task yield the data layout in Table 30.6.

Thus, we have $T/30 + T/22 = 3/5$, which yields $T = 7.62$ days. Consequently, the completion time of the task is $(12 + T) = 19.62$ days from time zero. The results of this example are summarized in the resource

TABLE 30.4 Data for Example 1

Resource Type i	Work Rate, r_i	Time, t_i	Work Done, w_i
RES 1	1/50	15	15/50
RES 2	r_2	15	$15(r_2)$
		Total	4/5

TABLE 30.5 Multiple Work Rate and Cost Data

Resource, i	Work Rate, r_i	Time, t_i	Work Done, w	Pay Rate, p_i	Total Cost, C_i
Machine A	r_1	t_1	$(r_1)(t_1)$	p_1	C_1
Machine B	r_2	t_2	$(r_2)(t_2)$	p_2	C_2
…	…	…	…	…	…
Machine n	r_n	t_n	$(r_n)(t_n)$	p_n	C_n
		Total	1.0		Budget

TABLE 30.6 Data for Example 2

Resource Type, i	Work Rate, r_i	Time, t_i	Work Done, w_i
RES 1	1/30	T	$T/30$
RES 2	1/22	T	$T/22$
		Total	3/5

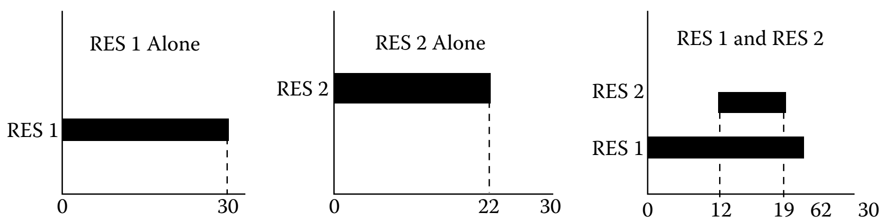


FIGURE 30.9 Resource schedule charts for RES 1 and RES 2.

schedule charts in Figure 30.9. It is assumed that both resources produce identical quality of work and that the respective work rates remain consistent. As mentioned earlier, the respective costs of the different types may be incorporated into the work rate analysis.

30.8 Resource Assignment Problem

Operations research techniques are frequently used to enhance resource allocation decisions [7, 8, 11, 12, 14, 15, 16]. One common resource allocation tool is the resource assignment algorithm. This algorithm can be used to enhance resource allocation decisions. Suppose that there are n tasks which must be performed by n workers. The cost of worker i performing task j is c_{ij}. It is desired to assign workers to the tasks in a fashion that minimizes the cost of completing the tasks. This problem scenario is referred to as the *assignment problem*. The technique for finding the optimal solution to the problem is called the *assignment method*. Like the transportation method, the assignment method is an iterative procedure that arrives at the optimal solution by improving on a trial solution at each stage of the procedure.

30.9 Resource Assignment Modeling

CPM and PERT can be used in controlling projects to ensure that the project will be completed on time. As was mentioned previously, these two techniques do not consider the assignment of resources to the tasks that make up a project. The *assignment method* can be used to achieve an optimal assignment of resources to specific tasks in a project. Although the assignment method is cost-based, task duration can be incorporated into the modeling in terms of time–cost relationships. Of course, task precedence requirements and other scheduling of the tasks. The objective is to minimize the total cost (TC). Thus, the formulation of the assignment problem is as follows:

Let

x_{ij} = 1 if worker i is assigned to task j, i, j = 1.2, …, n.

x_{ij} = 0 if worker i is not assigned to task j.

c_{ij} = cost of worker i performing task j.

$$\text{Minimize: } z = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij}$$

$$\text{Subject to: } \sum_{j=1}^{n} x_{ij} = 1, i = 1, 2, \ldots, n$$

$$\sum_{i=1}^{n} x_{ij} = 1, j = 1, 2, \ldots, n$$

$$x_{ij} \geq 0,\ i,j = 1,2,\ldots,n$$

It can be seen that the above formulation is a transportation problem with $m=n$ and all supplies and demands are equal to one. Note that we have used the nonnegativity constraint, $x_{ij} \geq 0$, instead of the integer constraint, $x_{ij} = 0$ or 1. However, the solution of the model will still be integer-valued. Hence, the assignment problem is a special case of the transportation problem with $m=n$, $S_i = 1$ (supplies), and $D_i = 1$ (demands). Conversely, the transportation problem can also be viewed as a special case of the assignment problem. A transportation problem can be modeled as an assignment problem and vice versa. The basic requirements of an assignment problem are as follows:

1. There must be two or more tasks to be completed.
2. There must be two or more resources that can be assigned to the tasks.
3. The cost of using any of the resources to perform any of the tasks must be known.
4. Each resource is to be assigned to one and only one task.

If the number of tasks to be performed is greater than the number of workers available, we will need to add *dummy workers* to balance the problem. Similarly, if the number of workers is greater than the number of tasks, we will need to add *dummy tasks* to balance the problem. If there is no problem of overlapping, a worker's time may be split into segments so that the worker can be assigned more than one task. In this case, each segment of the worker's time will be modeled as a separate resource in the assignment problem. Thus, the assignment problem can be extended to consider partial allocation of resource units to multiple tasks.

Although the assignment problem can be formulated for and solved by the simplex method or the transportation method, a more efficient algorithm is available specifically for the assignment problem. The method, known as the *Hungarian method*, is a simple iterative technique. Details of the assignment problem and its solution techniques can be found in operations research texts. As an example, suppose that five workers are to be assigned to five tasks on the basis of the cost matrix presented in Table 30.7. Task 3 is a machine-controlled task with a fixed cost of $800 regardless of which worker it is assigned to. Using the assignment, we obtain the optimal solution presented in Table 30.8, which indicates the following:

$$x_{15}=1,\ x_{23}=1,\ x_{31}=1,\ x_{44}=1, \text{ and } x_{52}=1$$

Thus, the minimum TC is given by

$$\text{TC}=c_{15}+c_{23}+c_{31}+c_{44}+c_{52}=\$(400+800+300+400+350)=\$2250$$

The technique of work rate analysis can be used to determine the cost elements that go into an assignment problem. The solution of the assignment problem can then be combined with the technique of CRD. This combination of tools and techniques can help enhance resource management decisions.

30.10 Line Balancing

Line balancing involves adjusting tasks at work stations so that it takes about the same amount of time to complete the work at each station. Most production facilities involve an integrated collection of work stations. Line balancing helps control the output rates in continuous production systems. As work is completed on a product at one station, it is passed on to the next for further processing. Cycle time is the time the product spends at each work station. The cycle time is dependent on the expected output from the line. The cycle time can be calculated based on the production rate. A balanced line refers to the equality of output of each successive work station on the assembly line. The maximum output of the line is controlled by its slowest operation. Perfect balance exists when each work station requires the same amount of time to perform its assigned tasks and there is no idle time at any of the work stations. Because of bottleneck operations and different work rates, perfect balance is rarely achieved.

The CRD approach can be effective in line balancing analysis. The work rate table can identify specific work rates that may need to be adjusted to achieve line balance within a specified margin. The *margin*, in

TABLE 30.7 Cost Matrix for Resource Assignment Problem

Worker	Task 1	Task 2	Task 3	Task 4	Task 5
1	300	200	800	500	400
2	500	700	800	1250	700
3	300	900	800	1000	600
4	400	300	800	400	400
5	700	350	800	700	900

TABLE 30.8 Solution to Resource Assignment Problem

Worker	Task 1	Task 2	Task 3	Task 4	Task 5
1	0	0	0	0	1
2	0	0	1	0	0
3	1	0	0	0	0
4	0	0	0	1	0
5	0	1	0	0	0

this case, is defined as the time deviation from perfect balance. The following definitions are important for line balancing analysis:

(1) *Work station*. The immediate work environment of an operator, which offers all the tools necessary to perform the tasks assigned to the operator.
(2) *Work element*. This is the smallest unit of work that cannot be divided between two or more operators without having schedule or resource conflict.
(3) *Operation*. A set of work elements assigned to a single work station.
(4) *Cycle time*. The time one unit of the product spends at each work station.
(5) *Balance delay*. The total idle time in the line as a result of unequal division of work between work stations.

Line balancing involves grouping work elements into stations so as to minimize the number of stations based on a desired output rate or cycle time. A precedence diagram is often used to determine the logical basis for grouping work elements. The CRD can be used for this purpose. Since there will be a large number of possible ways to group work elements, the analysis must use precedence constraints to eliminate the infeasible groupings.

A conveyor system is the simplest type of line balancing. The conveyor moves work elements past work stations at a constant rate. Each station is expected to perform its task within the time period allowed by the conveyor speed. A sign of imbalance occurs when work piles up in a particular station because more units of the product arrive at the station before the preceding ones are dispatched. In this case, the next operator in the line will experience some idle time while waiting for work to arrive from the preceding station.

The choice of cycle time depends on the desired output rate. The minimum number of work stations is the total work element duration divided by the cycle time. Fractional results are rounded up to the next higher integer value. This is calculated as:

$$n = \frac{\sum_{i=1}^{k} t_i}{C}$$

where k = number of work elements; t_i = processing time of work element i; C = cycle time.

In most cases, the theoretical minimum number of work stations will be impossible to achieve because of physical constraints in the production line and/or work element times that are not compatible. It should be noted that the theoretical minimum number of stations is not necessarily devoid of idle times. A balance with the theoretical minimum number of work stations will contain the least total idle time. The maximum possible efficiency with the theoretical minimum number of stations is represented as:

$$f_{\max} = \frac{\sum_{i=1}^{k} t_i}{nC}$$

For cases where the theoretical minimum number of stations cannot be achieved, the actual efficiency is computed by substituting the actual number of stations, m, for n in the above equation. Since actual efficiency, f_a, will be less than or equal to $f_{\max}$, the analyst would attempt to increase efficiency toward the maximum value by rearranging work elements. Several mathematical and heuristic methods are used for investigating the rearrangements. These include linear programming, dynamic programming, computer simulation, trial-and-error, and the ranked positional weight technique.

30.11 Human Resource Management

Human resources make projects successful. Human resources are distinguished from other resources because of the ability to learn, adapt to new project situations, and set goals. Human resources, technology resources, and management resources must coexist to pursue project goals. Managing human resources involves placing the right people with the right skills in the right jobs in the right environment. Good human resource management motivates workers to perform better. Both individual and organizational improvements are needed to improve overall quality by enriching jobs with the following strategies:

- Specify project goals in unambiguous terms
- Encourage and reward creativity on the job
- Eliminate mundane job control processes
- Increase accountability and responsibility for project results
- Define jobs in terms of manageable work packages that help identify line of responsibility
- Grant formal authority to make decisions at the task level
- Create advancement opportunities in each job
- Give challenging assignments that enable a worker to demonstrate his/her skill
- Encourage upward (vertical) communication of ideas
- Provide training and tools needed to get job done
- Maintain a stable management team

Several management approaches are used to manage human resources. Some of these approaches are formulated as direct responses to the cultural, social, family, or religious needs of workers. Examples of these approaches are:

- Flextime
- Religious holidays
- Half-time employment

These approaches can have a combination of several advantages. Some of the advantages are for the employer, while some are for the workers. The advantages are presented below:

- Low cost
- Cost savings on personnel benefits
- Higher employee productivity
- Less absenteeism
- Less work stress
- Better family/domestic situation, which may have positive effects on productivity

Work force retraining is important for automation projects. Continuing education programs should be developed to retrain people who are only qualified to do jobs that do not require skilled manpower. The retraining will create a ready pool of human resource that can help boost manufacturing output and competitiveness. Management stability is needed to encourage workers to adapt to the changes in industry. If management changes too often, workers may not develop a sense of commitment to the policies of management.

The major resource in any organization is manpower both technical and nontechnical. People are the overriding factor in any project life cycle. Even in automated operations, the role played by whatever few people are involved can be very significant. Such operations invariably require the services of technical people with special managerial and professional needs. The high tech manager in such situations would need special skills in order to discharge the managerial duties effectively. The manager must have auto-management skills that relate to the following:

- Managing self
- Being managed
- Managing others

Many of the managers who supervise technical people rise to the managerial posts from technical positions. Consequently, they often lack the managerial competence needed for the higher offices. In some cases, technical professionals are promoted to managerial levels and then transferred to administrative posts in functional areas different from their areas of technical competence. The poor managerial performance of these technical managers is not necessarily a reflection of poor managerial competence, but rather an indication of the lack of knowledge of the work elements in their surrogate function. Any technical training without some management exposure is, in effect, an incomplete education. Technical professionals should be trained for the eventualities of their professions.

In the transition from the technical to the management level, an individual's attention would shift from detail to overview, specific to general, and technical to administrative. Since most managerial positions are earned based on qualifications (except in aristocratic and autocratic systems), it is important to train technical professionals for possible administrative jobs. It is the responsibilities of the individual and the training institution to map out career goals and paths and institute specific education aimed at the realization of those goals. One such path is outlined below:

(1) *Technical professional*. This is an individual with practical and technical training and/or experience in a given field; such as industrial engineering. The individual must keep current in his/her area of specialization through continuing education courses, seminars, conferences, and so on. The mentor program, which is now used in many large organizations, can be effectively utilized at this stage of the career ladder.
(2) *Project manager*. This is an individual assigned the direct responsibility of supervising a given project through the phases of planning, organizing, scheduling, monitoring, and control. The managerial assignment may be limited to just a specific project. At the conclusion of the project, the individual returns to his/her regular technical duties. However, his/her performance on the project may help identify him/her as a suitable candidate for permanent managerial assignment later on.
(3) *Group manager*. This is an individual who is assigned direct responsibility to plan, organize, and direct the activities of a group of people with a specific responsibility—for example, a computer data security advisory committee. This is an ongoing responsibility that may repeatedly require the managerial skills of the individual.
(4) *Director*. An individual who oversees a particular function of the organization. For example, a marketing director has the responsibility of developing and implementing the strategy for getting the organization's products to the right market, at the right time, at the appropriate price, and in the proper quantity. This is a critical responsibility that may directly affect the survival of the organization. Only the individuals who have successfully proven themselves at the earlier career stages get the opportunity to advance to the director's level.
(5) *Administrative manager*. This is an individual who oversees the administrative functions and staff of the organization. His/her responsibilities cut across several functional areas. He/she must have proven his/her managerial skills and diversity in previous assignment.

The above is just one of the several possible paths that can be charted for a technical professional as he/she gradually makes the transition from the technical ranks to the management level. To function effectively, a manager must acquire nontechnical background in various subjects. His/her experience, attitude, personality, and training will determine his/her managerial style. His/her appreciation of the human and professional needs of his subordinates will substantially enhance his/her managerial performance. Examples of subject areas in which a manager or an aspiring manager should get training include the ones outlined below:

(1) Project management
    (a) *Scheduling and budgeting*: Knowledge of project planning, organizing, scheduling, monitoring, and controlling under resource and budget restrictions.
    (b) *Supervision*: Skill in planning, directing, and controlling the activities of subordinates.
    (c) *Communication*: Skill of relating to others both within and outside the organization. This includes written and oral communication skills.

(2) Personal and personnel management
  (a) *Professional development*: Leadership roles played by participating in professional societies and peer recognition acquired through professional services.
  (b) *Personnel development*: Skills needed to foster cooperation and encouragement of staff with respect to success, growth, and career advancement.
  (c) *Performance evaluation*: Development of techniques for measuring, evaluating, and improving employee performance.
  (d) *Time management*: Ability to prioritize and delegate activities as appropriate to maximize accomplishments within given time periods.

(3) Operations management
  (a) *Marketing*: Skills useful for winning new business for the organization or preserving existing market shares.
  (b) *Negotiating*: Skills for moderating personnel issues, representing the organization in external negotiations, or administering company policies.
  (c) *Estimating and budgeting*: Skills needed to develop reasonable cost estimates for company activities and the assignment of adequate resources to operations.
  (d) *Cash flow analysis*: An appreciation for the time value of money, manipulations of equity and borrowed capitals, stable balance between revenues and expenditures, and maximization of returns on investments.
  (e) *Decision analysis*: Ability to choose the direction of work by analyzing feasible alternatives.

A technical manager can develop the above skills through formal college courses, seminars, workshops, short courses, professional conferences, or in-plant company training. Several companies appreciate the need for these skills and are willing to bear the cost of furnishing their employees with the means of acquiring the skills. Many of the companies have custom formal courses which they contract out to colleges to teach for their employees. This is a unique opportunity for technical professionals to acquire managerial skills needed to move up the company ladder.

Technical people have special needs. Unfortunately, some of these needs are often not recognized by peers, superiors, or subordinates. Inexperienced managers are particularly prone to the mistake of not distinguishing between technical and nontechnical professional needs. In order to perform more effectively, a manager must be administratively adaptive. He/she must understand the unique expectations of technical professionals in terms of professional preservation, professional peers, work content, hierarchy of needs, and the technical competence or background of their managers.

30.11.1 Professional Preservation

Professional preservation refers to the desire of a technical professional to preserve his/her identification with a particular job function. In many situations, the preservation is not possible due to a lack of manpower to fill specific job slots. It is common to find people trained in one technical field holding assignments in other fields. An incompatible job function can easily become the basis for insubordination, egotism, and rebellious attitudes. While it is realized that in any job environment there will sometimes be the need to work outside one's profession, every effort should be made to match the surrogate profession as close as possible. This is primarily the responsibility of the human resources manager.

After a personnel team has been selected in the best possible manner, a critical study of the job assignments should be made. Even between two dissimilar professions, there may be specific job functions that are compatible. These should be identified and used in the process of personnel assignment. In fact, the mapping of job functions needed for an operation can serve as the basis for selecting a project team. In order to preserve the professional background of technical workers, their individualism must be understood. In most technical training programs, the professional is taught how to operate in the following ways:

(1) Make decisions based on the assumption of certainty of information.
(2) Develop abstract models to study the problem being addressed.
(3) Work on tasks or assignments individually.
(4) Quantify outcomes.
(5) Pay attention to exacting details.
(6) Think autonomously.
(7) Generate creative insights to problems.
(8) Analyze systems "operatability" rather than profitability.

However, in the business environment, not all of the above characteristics are desirable or even possible. For example, many business decisions are made with incomplete data. In many situations, it is unprofitable to expend the time and efforts to seek perfect data. As another example, many operating procedures are guided by company policies rather than creative choices of employees. An effective manager should be able to spot cases where a technical employee may be given room to practice his professional training. The job design should be such that the employee can address problems in a manner compatible with his professional training.

30.11.2 Professional Peers

In addition to having professionally compatible job functions, technical people like to have other project team members to whom they can relate technically. A project team consisting of members from diversely unrelated technical fields can be a source of miscommunication, imposition, or introversion. The lack of a professional associate on the same project can cause a technical person to exhibit one or more of the following attitudes:

(1) Withdraw into a shell; and contribute very little to the project by holding back ideas that he/she feels the other project members cannot appreciate.
(2) Exhibit technical snobbery; and hold the impression that only he/she has the know-how for certain problems.
(3) Straddle the fence on critical issues; and develop no strong conviction for project decisions.

Providing an avenue for a technical "buddy system" to operate in an organization can be very instrumental in ensuring congeniality of personnel teams and in facilitating the eventual success of project endeavors. The manager in conjunction with the selection committee (if one is used) must carefully consider the mix of the personnel team on a given project. If it is not possible or desirable to have more than one person from the same technical area on the project, an effort should be made to provide as good a mix as possible. It is undesirable to have several people from the same department taking issues against the views of a lone project member from a rival department. Whether it is realized or not, whether it is admitted or not, there is a keen sense of rivalry among technical fields. Even within the same field, there are subtle rivalries between specific functions. It is important not to let these differences carry over to a project environment.

30.11.3 Work Content

With the advent of new technology, the elements of a project task will need to be designed to take advantage of new developments. Technical professionals have a sense of achievement relative to their expected job functions. They will not be satisfied with mundane project assignments that will bring forth their technical competence. They prefer to claim contribution mostly where technical contribution can be identified. The project manager will need to ensure that the technical people of a project have assignments for which their background is really needed, It will be counterproductive to select a technical professional for a project mainly on the basis of personality. An objective selection and appropriate assignment of tasks will alleviate potential motivational problems that could develop later in the project.

30.11.4 Hierarchy of Needs

Recalling Maslow's hierarchy of needs, the needs of a technical professional should be more critically analyzed. Being professionals, technical people are more likely to be higher up in the needs hierarchy. Most of their basic necessities for a good life would already have been met. Their prevailing needs will tend to involve esteem and self-actualization. As a result, by serving on a project team, a technical professional may have expectations that cannot usually be quantified in monetary terms. This is in contrast to nontechnical people who may look forward to overtime pay or other monetary gains that may result from being on the project. Technical professionals will generally look forward to one or several of the following opportunities.

(1) *Professional growth and advancement.* Professional growth is a primary pursuit of most technical people. For example, a computer professional has to be frequently exposed to challenging situations that introduce new technology developments and enable him to keep abreast of his field. Even occasional drifts from the field may lead to the fear of not keeping up and being left behind. The project environment must be reassuring to the technical people with regard to the opportunities for professional growth in terms of developing new skills and abilities.
(2) *Technical freedom.* Technical freedom, to the extent permissible within the organization, is essential for the full utilization of a technical background. A technical professional will expect to have the liberty of determining how best the objective of his assignment can be accomplished. One should never impose a work method on a technical professional with the assurance that "this is the way it has always been done and will continue to be done!" If the worker's creative input to the project effort is not needed, then there is no need having him or her on the team in the first place.
(3) *Respect for personal qualities.* Technical people have profound personal feelings despite the mechanical or abstract nature of their job functions. They will expect to be respected for their personal qualities. In spite of frequently operating in professional isolation, they do engage in interpersonal activities. They want their nontechnical views and ideas to be recognized and evaluated based on merit. They don't want to be viewed as "all technical". An appreciation for their personal qualities gives them the sense of belonging and helps them to become productive members of a project team.
(4) *Respect for professional qualification.* A professional qualification usually takes several years to achieve and is not likely to be compromised by any technical professional. Technical professionals cherish the attention they receive due to their technical background. They expect certain preferential treatments. They like to make meaningful contributions to the decision process. They take approval of their technical approaches for granted. They believe they are on a project because they are qualified to be there. The project manager should recognize these situations and avoid the bias of viewing the technical person as being conceited.
(5) *Increased recognition.* Increased recognition is expected as a by-product of a project effort. The technical professional, consciously or subconsciously, views his participation in a project as a means of satisfying one of his higher-level needs. He/she expects to be praised for the success of his/her efforts. He/she looks forward to being invited for subsequent technical endeavors. He/she savors hearing the importance of his contribution being related to his/her peers. Without going to the extreme, the project manager can ensure the realization of the above needs through careful comments.

(6) *New and rewarding professional relationship.* New and rewarding professional relationships can serve as a bonus for a project effort. Most technical developments result from joint efforts of people that share closely allied interests. Professional allies are most easily found through project groups. A true technical professional will expect to meet new people with whom he/she can exchange views, ideas, and information later on. The project atmosphere should, as a result, be designed to be conducive to professional interactions.

30.11.5 Quality of Leadership

The professional background of the project leader should be such that he/she commands the respect of technical subordinates. The leader must be reasonably conversant with the base technologies involved in the project. He/she must be able to converse intelligently on the terminologies of the project topic and be able to convey the project ideas to upper management. This serves to give him/her technical credibility. If technical credibility is lacking, the technical professionals on the project might view him/her as an ineffective leader. They will consider it impossible to serve under a manager to whom they cannot relate technically.

In addition to technical credibility, the manager must also possess administrative credibility. There are routine administrative matters that are needed to ensure a smooth progress for the project. Technical professionals will prefer to have those administrative issues successfully resolved by the project leader so that they can concentrate their efforts on the technical aspects. The essential elements of managing a group of technical professionals involve identifying the unique characteristics and needs of the group and then developing the means of satisfying those unique needs.

Recognizing the peculiar characteristics of technical professionals is one of the first steps in simplifying project management functions. The nature of manufacturing and automation projects calls for the involvement of technical human resources. Every manager must appreciate the fact that the cooperation or the lack of cooperation from technical professionals can have a significant effect on the overall management process. The success of a project can be enhanced or impeded by the management style utilized.

30.11.6 Work Simplification

Work simplification is the systematic investigation and analysis of planned and existing work systems and methods for the purpose of developing easier, quicker, less fatiguing, and more economic ways of generating high-quality goods and services. Work simplification facilitates the content of workers, which invariably leads to better performance. Consideration must be given to improving the product or service, raw materials and supplies, the sequence of operations, tools, work place, equipment, and hand and body motions. Work simplification analysis helps in defining, analyzing, and documenting work methods.

30.12 Alternate Optimization Formulations for CRD

The sections that follow present alternate optimization formulations for the CRD method. Each formulation can be adapted and enhanced for specific problem scenario of interest.

30.12.1 Mixed-integer Programming Formulation

Generally, heuristic procedures have been asserted to be expedient for scheduling large and complex projects which occur in practice due to its few computational requirements and its minimal computational time. However, heuristic methods do not always produce stable or exact schedules. In proffering solutions to the resource constrained scheduling problem therefore, there is a need to develop a method that provides an optimal or near-optimal schedule of the limited resources of a project such that the sequence in use, and the availability of such resources are not violated throughout the entire duration of the project.

As an extension of the approach proposed by Pritsker et al. [12] in addressing multiproject scheduling with limited resources, a zero–one mathematical programming model is developed to improve the practicality of the CRD. This model can be used to analyze interdependencies, criticality, substitution, conflicting resource priorities and variations in the availability of resources.

In the next section, notations used and the formulation developed for the resource scheduling problem were introduced.

30.12.1.1 Definition of Problem Variables and Integer Problem Formulation

Table 30.9 presents the notation used in this chapter. The model supposes that the variables J, M, d_{rj}, a_r and a_{rt} are integer valued and known with certainty. In addition, d_{rj}, ≥ 0, $a_r \geq 0$, $a_{rt} \geq 0$, and $d_{mj} \geq 0$ for all $j = 1, \ldots, n\}$and all r = $\{1, \ldots, m\}$.

30.12.2 Problem Formulation

In formulating the IP problem, this chapter focuses on a single project where preemption or splitting of jobs was not allowed (once a resource r is scheduled for use, it must work on the assigned job j until the job is completed in d_{rj} time units) and where resources utilized are discrete in nature.

TABLE 30.9 Notations used to Describe the Integer Programming Model

Symbol	Definition
n	Number of jobs
J	Set of all jobs; $J = \{1, \ldots\ldots, n\}$
j	Index for the jobs: $j = 1, \ldots\ldots, n$
d_j	Duration of job j
d_{rj}	Duration of use of resource r by job j
d_r	Duration of use of resource r at a given time
P_r/F_r	Det of jobs which immediately precede/follow job
P_r^*/F_r^*	Set of jobs which precede/follow job j
C	Completion time of the project
$\bar{T}$	End of planning horizon
t	Index for periods; $t = 1, \ldots\ldots, T$
M	Number of renewable resource types
R	Set of all renewable resource types; $R = \{1, \ldots\ldots, m\}$with resource m being the unique terminal resource without successors
r	Index for the renewable resource types; $r = \{1, \ldots\ldots, m\}$
ES_r	Earliest possible period by which resource r can be started
LS_r	Latest possible period by which resource r can be started
EF_r	Earliest possible period by which resource r can be released
LF_r	Latest possible period by which resource r can be released
a_r	Constant per period availability of resource type r for all t
a_{rt}	Availability of resource type r in period t
u_{rj}	Per period usage of resource type r by job j

For this resource scheduling problem, we seek to minimize the time by which all resources being utilized in the project are released. But first we define the following zero–one variables.

$$x_{rt} = \begin{cases} 1, \text{ if resource } r \text{ is being released in period } t \\ 0, \text{ otherwise} \end{cases}$$

$$x_t = \begin{cases} 1, \text{ if project is completed by period } t \\ 0, \text{ otherwise} \end{cases}$$

We also define the period SF_r in which a resource r is scheduled to be released as;

$$SF_r = \sum_{t=EF_r}^{LF_r} t.x_{rt}$$

Since resources must be scheduled in a manner that optimizes our objective, we define an objective function given by Equation 30.1.

Objective function:

$$\text{Minimize } CT = \sum_{t=EF_m}^{LF_m} t.x_{mt} \tag{30.1}$$

where EF_m and LF_m are the earliest and latest release times of the last resource m being used in completing the project.

This objective function is related to requirements and limitations as follows:

30.12.2.1 Finish Time Constraints

(i) *Each resource has exactly one release time.*

$$\sum_{t=EF_r}^{LF_r} x_{rt} = 1, \quad \text{for } r \in R\,\theta \tag{30.2}$$

Since the value of any one x_{rt} can be determined by the values of the others in this constraint, Equation 30.2 can be stated in a better form as

$$\sum_{t=EF_r}^{LF_r-1} x_{rt} \leq 1 \tag{30.3}$$

and define $x_{rt}(LF_r) \equiv 1 - \sum_{t=EF_r}^{LF_r-1} x_{rt}$.

(ii) *The project has exactly one completion time.*

The project cannot be completed by period t until $\sum_{q=EF_r}^{t-1} x_{rt} = 1$ for all $M \in r$ resources being used for the completion of the project such that:

$$x_t \leq (1/m) \sum_{m=1}^{M} \sum_{q=EF_r}^{t-1} x_{rt} \quad \left(\text{for } r = 1, 2, \ldots, m;\ t = c, c+1, \ldots, \overline{T}\right) \tag{30.4}$$

30.12.2.2 Precedence Constraint

A precedence constraint is needed to express the limitation that is involved when a resource cannot be started until one or more other resources have been released. For the purpose of formulating the equation for this constraint, we might consider a case where a resource s must precede a resource q.

If t_s and t_q denote the release periods of resources s and q respectively, then:

$$t_s + d_q \le t_q$$

The precedence constraint can then be written as:

(i)
$$\sum_{t=EFq}^{LFq} (t - d_q).x_{qt} - \sum_{t=EF_s}^{LF_s} x_{st} \ge 0 \quad \text{for } q, s \in R \text{ and } s \in p_q \tag{30.5}$$

Or

(ii)
$$\sum_{t=EFs}^{LF_s} t x_{st} + d_q \le \sum_{t=EF_q}^{LF_q} x_{qt} \quad \text{for } q, s \in R \text{ and } s \in p_q \tag{30.6}$$

Since the exact release time of resource s and q lies in the time-window [EF_s, LS_s; EF_q, LS_q], both Equations 30.4 and 30.5 implies that the time lag between the start time of resource q and the completion time of resource s is an amount ≥ 0.

30.12.2.3 Resource Availability Constraint

In order to ensure that the usage of any one resource by the jobs do not exceed the availability of the resource at any given period, we have to consider the utilization of the resources by all the jobs that are being processed within that period. A resource is being utilized by job j in period t if the job is completed and/or the resource is released in period f where $t \le f \le t + d_{rj} - 1$. Therefore the resource constraint can be written as:

$$\sum_{r=1}^{M} \sum_{f=t}^{t+d_{rj}-1} u_{rj} x_{rt} \le a_{rt} \quad \left(t = \min\ ES_r, \ldots, \max \overline{T}\right) \tag{30.7}$$

For this constraint to hold, we have to establish predetermined values of x_{rt} namely:

$$(x_{rt} \equiv 0 \text{ for } t < EF_r \text{ and } x_{rt} \equiv 0 \text{ for } t > LF_r).$$

30.12.2.4 Concurrency and Nonconcurrency of Resources

For most projects, a resource r is utilized only once by a job j for a period d_{rj}. However, it is possible for a resource to be utilized more than once by two or more different jobs from the set of n jobs that constitutes the project. If a resource $h \in R$ appears more than once in the course of the project, it is treated as another type of resource (such as $k \in R$) at each of the appearance after its first one, such that the availability per period at each appearance of this resource is split among its total number of appearances. A concurrency constraint on resources s and q is necessary if they must be utilized simultaneously. It can be obtained by requiring $x_{st} = x_{qt}$ or by combining their availabilities and treating s and q as a single resource. A nonconcurrency constraint on resources s and q is necessary if they must not be performed simultaneously.

30.12.2.5 Substitutability Constraints

In certain situations commonly experienced in practice, it is possible to use alternative resources to perform a set of jobs such that there is a difference in job durations when a particular job of this set is performed by different resources. For example, a resource of higher skill r' may accomplish a job j' in $d_{r'j'}$ time units while a resource of lower skill r'' accomplishes the same job in $d_{r''j'}$ time units. In dealing with this condition, we define the set of mutually exclusive resources one of which must be utilized such that $LF_{r'j'} = LF_{r''j'} = LF_{rj'}$. The required release times of either of these resources, but not both, must be any time before the period $LF_{rj'}$. This is represented by the constraint equation:

$$\sum_{q=\min(EF_{r'j'}, EF_{r''j'})}^{LF_{rj'}} \left(x_{r'q} + x_{r''q}\right) = 1 \tag{30.8}$$

The modified project completion constraint corresponding to Equation 30.4 would then be:

$$x_t \le (1/M)\left[\sum_{m=1\cdot m\ne r'}^{M} \sum_{q=EF_r}^{t-1} x_{rt} + \sum_{q=\min(EF_{r'j'}, EF_{r''j'})}^{LF_{rj'}} \left(x_{r'q} + x_{r''q}\right)\right] \tag{30.9}$$

30.13 Conclusion

The techniques presented in this chapter have been applied effectively to practical projects. The provide more realistic handling of project scheduling beyond the conventional PERT/CPM network analysis. The conventional CPM and PERT approaches assume unlimited resource availability in project network analysis. In realistic projects, both the time and resource requirements of activities should be considered in developing network schedules. Projects are subject to three major constraints: time limitations, resource constraints, and performance requirements. Since these constraints are difficult to satisfy simultaneously, trade-offs must be made. The smaller the resource base, the longer the project schedule. The CRD and RS charts are simple extensions of very familiar management tools. They are simple to use, and they convey resource information quickly. They can be used to complement existing resource management tools. CRD can be modified for specific resource planning, scheduling, and control needs in both small and large projects.

References

1. Ashok Kumar, V.K. and L.S. Ganesh. 1998. Use of petri nets for resource allocation in projects. *IEEE Transactions on Engineering Management*, 45(1), 49–56.
2. Badiru, A.B. 1992. Critical resource diagram: A new tool for resource management. *Industrial Engineering,* 24(10), 58–59, 65.
3. Badiru, A.B. 1993. Activity resource assignment using critical resource diagramming. *Project Management Journal,* 24(3), 15–21.
4. Badiru, A.B. 1995. Incorporating learning curve effects into critical resource diagramming. *Project Management Journal,* 26(2), 38–45.
5. Campbell, G.M. 1999. Cross-utilization of workers whose capabilities differ. *Management Science,* 45(5), 722–32.
6. Carraway, R.L. and R.L. Schmidt. 1991. An improved discrete dynamic programming algorithm for allocating resources among interdependent projects. *Management Science,* 37(9), 1195–200.
7. Cui, W., J. Qin, X. Cao, C. Yue, and Z. Yin. 2006. Critical resources identification in constant resource constrained projects. IEEE Conference on Systems, Man, and Cybernetics, October 8–11, Taipei, Taiwan.

8. Goldstein E.G. and S. Dempe. 2002. A minimax resource allocation problem with variable resources. *European Journal of Operational Research,* 136, 46–56.
9. Lai, K.K. and L. Li. 1999. A dynamic approach to multiple-objective resource allocation problem. *European Journal of Operational Research*, 117, 293–309.
10. Lova, A., C. Maroto, and P. Tormos. 2000. A multicriteria heuristic method to improve resource allocation in multiproject scheduling. *European Journal of Operational Research*, 127, 408–24.
11. Narasimhan, M. and J. Ramanujam. 2001. A fast approach to computing exact solutions to the resource constrained scheduling problem. *ACM Transactions on Design Automation of Electronic Systems*, 6(4), 490–500.
12. Pritsker, A.A.B., L.J. Watters, and P.M. Wolfe. 1969. Multiproject scheduling with limited resources: A zero-one programming approach. *Management Science,* 16(1), 93–108.
13. Stinson, J.P., E.W. Davis, and B.M. Khumawala. 1978. Mutiple resource-constrained scheduling using branch and bound. *AIIE Transactions,* 10(3), 252–59.
14. Talbot, F.B. 1982. Resource-constrained project scheduling with time resource tradeoffs: The non-preemptive case. *Management Science,* 28(10), 1197–210.
15. Yang, B.-H., A. Badiru, and S. Saripalli. 2005. A facility resource scheduling for two-stage supply chain. *Journal of the Korean Institute of Plant Engineering*, 10(1), 67–78.
16. Yang, B.-H., A. Badiru, and S. Saripalli. 2004. Resource scheduling for supply chain model using critical resource diagramming. Proceedings of the Asia Pacific Management Conference, South Korea, 1–12.

31

Innovative Techniques and Practical Software Tools for Addressing Military Analytical Problems

Bobbie L. Foote
United States Military Academy

Simon R. Goerger
National War College

31.1 Coping with Military Procurement Constraints Mandated by Congress as a Constraint on Inventory Models

One of the gaps between theory and practice is the system that executes the solution of an inventory model. We discuss here an example to show how one copes with this system in translating a dynamic programming inventory model that has (random) item demand, random lead-times, with order costs, holding costs, and shortage costs. This example comes from a USN case at Aviation Supply Office in Philadelphia. The main problem is coping with the requirement of competitive bidding and the time required for advertising, receiving bids, and contract negotiation and definition.

Two issues dominate: the determination of parameters for the model and actually executing the strategy. In this case the Navy records produced values for ordering costs, holding costs that were defensible and had a reasonable accuracy. Shortage costs were another matter. These parts were not critical in that a shortage simply lost part of the capability (i.e. a percent of the fleet down as opposed to the whole fleet). First, what is the cost of a fighter plane being down due to being a part short? If it is short only one part then a reasonable estimate is the purchase price of the plane divided by its lifetime in days. This produced a value that indicated too much inventory using the standard determination. This would exceed budget. A further question was: what if the plane was down due to two or more parts short. In this case if it is found that another part is short does the Navy incur an additional shortage cost? This means that a shortage cost is a function of the inventory levels of a whole group of parts. This problem could be dealt with by developing an event tree and costing out each endpoint. This was not DOD policy. The next issue is the execution. Ideally the Liberatore (1977) model is evaluated and the amount computed is ordered and the random lead-time is incurred. The randomness of the lead-time is due to these kinds of parts having special materials, therefore the factories having unpredictable bottlenecks due to the special processing involved. Further the scheduling of orders is not only dependent on due date but on system economics. A small order that was made in October may be scheduled because there is a small gap of time to produce it. An order received in January that is three times the October order may not be scheduled because there is not time enough between emergency orders to get it done. However, one more factor is involved. This is the legal requirement to open the purchase to competitive bid. This means working to address the requirement for set asides for small and minority owned business. This adds a cost and an uncertainty that violates the model requirements.

An approach was to analyze the model over sets of possible parameters to see if certain policies were optimal over a wide variety of sets. The thesis by Kebriae (1981) showed that this was true. The most likely policies were to order 3, 6, 9 or 12 months supply. In this case the uncertainty of the shortage value made it hard to pick one of the policies. Some hand calculations over a set of parts and combinations of parameter values made a 9 months order look like a good policy with a high degree of belief. In addition a paper by Foote (1972) showed in a dynamic programming model with random demand and constant lead-time that the infinite horizon model was a good approximation to the optimal policy if the number of periods was three or more. Using the infinite horizon model as a heuristic again showed a large order quantity for these parameters. These two heuristics gave quantities that showed a possible violation of maximum holding rules unless something could offset. One fact helped. The government had industrial engineers in the plants that verified company data. Past bids had up to a 100 pages of company parameters including learning data. From this data one could infer price breaks companies could give for being given a large order. These price breaks were due to the amortization of setup costs and the learning that occurs with large lots. The decision was made to put out for competitive bid to acquire 3, 6, 9 and 12 months expected supply. Each company bid a price per unit for each quantity. A standard engineering economy present value analysis using the government borrowing rate was then conducted using only cash flows that were not allocated accounting values such as holding and obsolescence costs and a bidder was selected without using shortage cost. Usually 9 or 12 months supply was selected. This gave an inventory level that was likely to be an optimal policy according to the Liberatore (1977) model. A later audit showed in hindsight that over a seven year period more than a billion dollars were saved.

This project showed the value of sensitivity analysis to show forms of optimal policies which inform the specific details of the request for proposal issued by the government.

Sometimes well intentioned regulations addressing real problems such as cronyism and noncompetition can lead to adverse cost consequences. Procurement regulations were originally set up to buy shoes, blankets, flour and other staples. One example is the requirement to bid each part purchased. However, if the parts all have a similar geometry and similar materials, then a group technology setup can develop huge savings. In these cases exceptions to the regulations must be sought. At Aviation Supply Office, ASO, an exception was sought for circular engine parts. A test contract was awarded with an inflation index. The large diverse order allowed an investment in group technology. The initial price was over 40% less than the last competitive bid history. Ten years later even with the inflation index parts still cost less in 1998 that in 1988. *Key principle: Regulations must harmonize with the cost structure of the industry which supplies produced items* (Irani 1999 and Bedworth et al. 1991).

31.1.1 Problems with Incommensurate Measures

Procurement models are particularly susceptible to problems of metrics. The classic newsboy problem where purchase costs are balanced against revenues is the classic where the measures commensurate. Both are real cash flows. The problem models are those that balance purchase costs, holding costs, shortage costs and "obsolescence costs". If holding costs are computed to be $15 dollars per item, one will not find a check written for $15. One may find a case where a real interest value was forfeited, but it will be less that $15. It is valid to use allocated costs for holding as there has to be recognition somehow that it costs to provide storage bins, MH equipment, labor to operate, utilities, data systems, etc. These systems once provided for are there and do not change in cash flow based on storage volume in general. These costs are better modeled in many cases as representing constraints. When storage volume exceeds capacity real cash flows occur. Similarly when goods deteriorate then real cash flows are forfeited when the material is not scrapped but held until rendered defective. Food rots, isotopes degrade, parts rust.

Similarly scrap models that consider obsolescence have definition problems for the cost of obsolescence. Modelers correctly gauge that as the design date of systems grows older then new systems with the same application are more effective and a value lost of some type will occur. At some point scrapping the older designs giving cash to buy more capable items makes sense. The problem is that these value changes are not linear. If a new design is not available, obsolescence "cost" is not accruing. In these cases it is better to model a lost value just like an interest payment foregone. A poor model can lead to items being scrapped and then a portion bought back at a premium which can in total be more than the total scrap sale. Part of the problem lies in exponential smoothing forecasting when the demand is lumpy. Cases occur where demand will occur every so often, say five years. The first demand may be 1000. Then there will be 0 demands for 60 months. The smoothed monthly forecast will be in effect a 0 monthly demand. The material is sold and immediately there will be a demand with some random demand with parameter 1000. Clearly scrap models must determine if a requirement still exists, if the government is required by treaty to supply it and if the demand is lumpy and what is the cause of the lumpy demand. Causes can be sudden war, scheduled exercises, etc. Instead of modeling an obsolescence cost, it would be better to consider a supply constraint for the cost model. One can simply sell off excess materiel unless storage is also a constraint. Figure 31.1 shows a simple event tree showing a way to calculate cost of a part request denied.

The shortage cost $S = P^{*}(0.0) + (1-P)^{*} n^{*}$ cost/day. This ignores that if the plane is already down it may be down later due to the part not being available. A more complex tree could then be defined.

31.2 The use of Fuzzy Arithmetic in Estimating the Probability of a Terrorist Attack

Fuzzy mathematics was meant to bring more rigor to the art of human judgment. We often make statements to police about the weight of a suspect. He weighs "about 180 lbs". The government states the terror threat is "high". There may be more precision in a witness's knowledge than just the most likely weight is 180. Usually witness estimates are multiples of five. They may see the person as chunky and would estimate low if they are wrong. Zadeh (1975) developed a subjective approach that enables a mathematical description. Let us look at Figure 31.2.

The figure illustrates that the highest degree of belief is at 180 but there is more leeway above than below 180. By using proper interrogation techniques one can establish end points and the maximum degree of belief (always less than 1). One sees here that the witness believes the weight cannot be less than 172 and not more than 195. The shape of the curve must be continuous but can be convex on the left and concave on the right. This adds somewhat more precision when a possible suspect is viewed by an officer. A "chunky" person is more likely to be vetted.

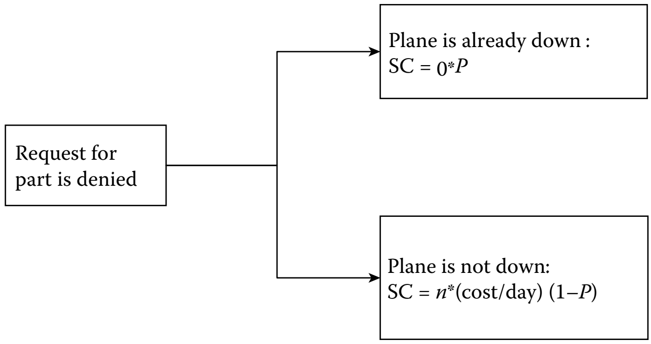


FIGURE 31.1 Illustration of cost of a part request denied event tree.

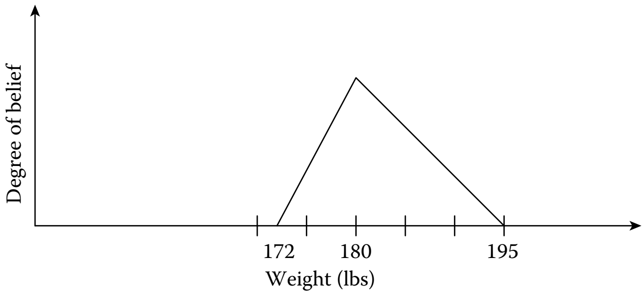


FIGURE 31.2 Illustration of a curve representing degree of belief.

In the hypothetical example that follows, we can add the concept of 'degree of belief' to the problem of calculating the probability of a terror attack in a specific locality such as New York City. We could just compute the average number of attacks over a set period of time, put this in the Poisson probability distribution and then compute the probability: $P = f(lambda, t)$ where *lambda* is the parameter of the Poisson which is the average number of events. But we have some ambiguity. If we choose different time periods, we will have different averages. Perhaps we can express the average as a fuzzy number.

Let us suppose we had a terror event at time 0, 86, 286, 586 and 635. This gives time periods of 86, 200, 300 and 49. This gives averages determined by 1/86, 2/286, 3/586, and 4/635. Since $t=635$ is now, we could have 1/49 as a monthly average. The biggest value is 1/49. Since this is nearest we use this as our highest degree of belief. We could give the degree of belief by a Delphi process as 0.4. The lowest value is 0.00512 and the highest is 1/49=0.0204 incidents per month. This would give rise to a natural triangular DB function. We can compute the curve by fitting the points (0.00512, 0) and (0.0204, 0.4). We can then compute the degree of belief in 0.01 incidents per month. The curve fit is DB=26.1780*(average)−0.13403 obtained by rounding to five places. The DB of 0.01 as a monthly average is 26.1780*0.01−0.13403 =0.1 27749 (Figure 31.3). We could try other degree of belief functions that are concave or convex. There is no evidence that using complex curves increases the utility of the process. The shape can be chosen by logic or by the Delphi process. One could choose a concave function if there is some specific terror threat that has been determined by intelligence services (Figure 31.4).

For a good and complete explanation of fuzzy logic and mathematics see McCauley-Bell and Crumpton-Young (2006).

One can now compute the probability of 0 terror events for the monthly average of 0.01 using the Poisson pdf as $e^{-0.01}/(0!) = 0.99$. The degree of belief of this probability is 0.1295. This can be repeated for

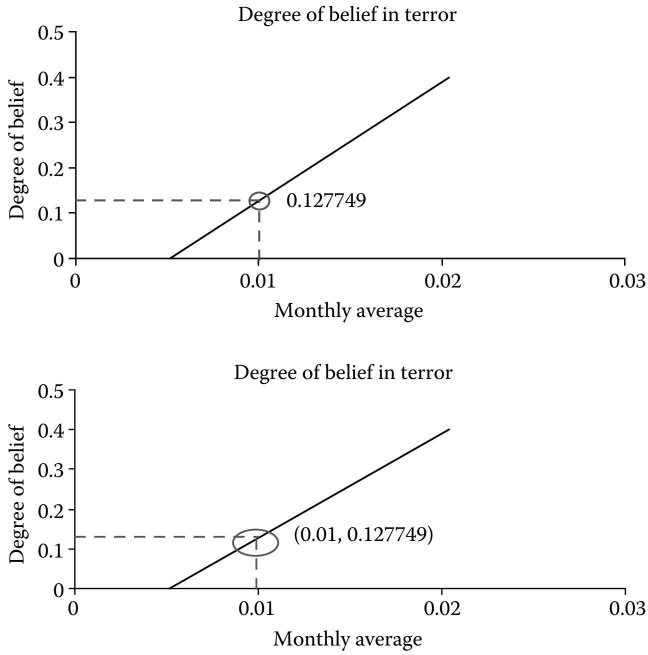


FIGURE 31.3 Illustration of result of Delphi process delineating degree of belief in the monthly average of terror attacks.

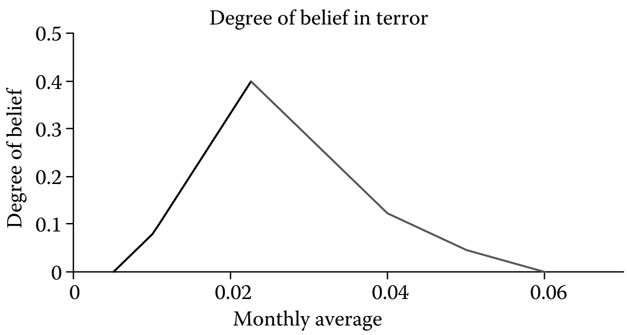


FIGURE 31.4 A concave degree of belief function that illustrates a high confidence in the mode and a rapid loss of confidence in other values.

each average that has a nonzero degree of belief and a curve representing each degree of belief for each probability of terrorist attack can be drawn. Similarly, one would have a degree of belief of 0.4 of a probability of 0.98 of no terrorist attack this month using the monthly average 0.0204 (Figure 31.5).

If our intelligence has a degree of belief that a terror event is feasible below 0.25, then we are on low alert as a population for clues. We are code green. If we have a higher DB but less that 0.5 we are at yellow. Everyone will review policy and recheck. If we are below 0.85 and above 0.5 we are at orange and will seek guidance, and above 0.85 we are red. If we are at red alert then businesses and schools will assign people to be full time vigilant, move parking away or block it, etc. and Homeland Security Agency (HSA) will advertise what to look for (Figure 31.6).

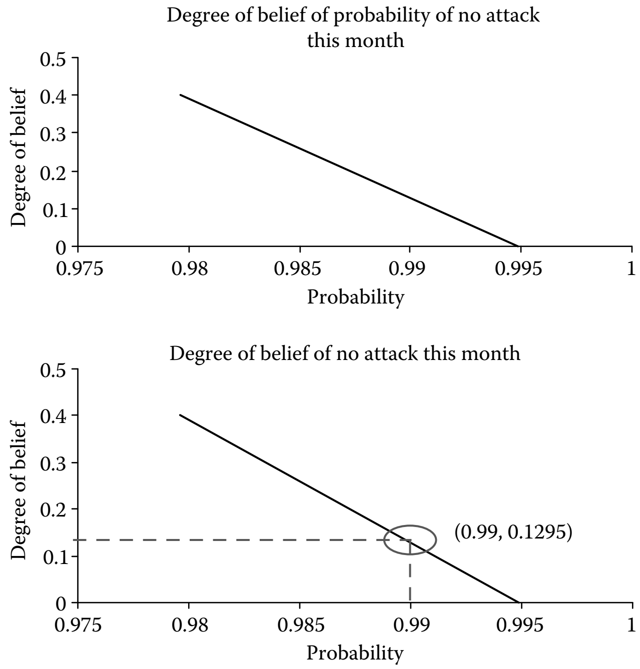


FIGURE 31.5 Degree of belief of probability of no attack.

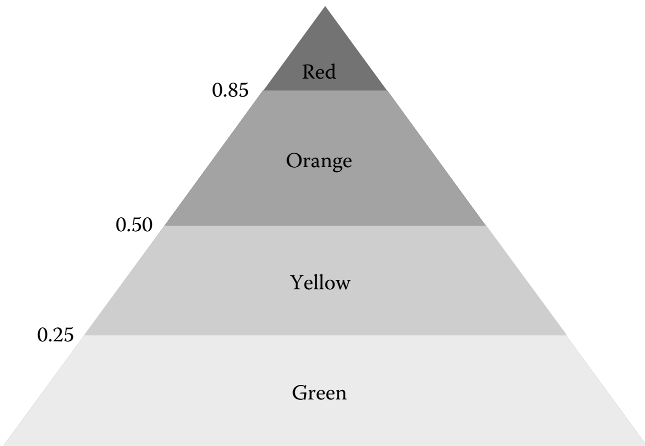


FIGURE 31.6 Color code of HSA warning.

31.3 Developing New Metrics to Measure Readiness for the Purposes of Modeling and Consequent Optimization

The standard measure of readiness for the Army currently is the percentage of equipment that is completely ready with no defective parts and the percentage of personnel that is up to date in all their required training. This system serves the Army well in peace time and has been effective. However, this set of measures ignores operational experience that shows that if some parts of equipment are defective that military value still exists and if training is not updated there is still military capability available from the personnel.

As an example, if an Apache helicopter cannot fire its Hellfire missiles, its 30 mm machinegun and its Hydra rockets, has no night vision nor a target acquisition system, it still can perform daylight surveillance. This has low value (utility) but it still can be valuable as a part of a specific mission. Clearly for a night attack there is zero value. But if it is known that the Apache can fly in daylight it can be used by a commander. The readiness of a helicopter can be defined as an illustrative example: R1, R2, R3, R4. This can be defined in nonmilitary terms as (1) ready for daylight flights; (2) can also fly at night; (3) can also target and fire one of its attack systems; (4) can also target and attack with its remaining systems. Let us illustrate with an idealized schematic of a helicopter represented by two tables. Notice if you have a top level system available then you cannot use it if you cannot fly. Thus there is a strong precedence constraint. All your prior capabilities must function for a given level capability to be expressed.

If we look at the two helicopters we see H2 has no rotor and both lack their hellfire capability (Figure 31.7). Clearly, if we have parts for the hellfire firing system (Kaczynski et al. 2003) we would put them in H1.

If H3 is available, it is clear that one would in a given circumstance invest in labor to remove parts 2 and 8 and put part 2 in H2 and in H1, part 8. We now have two C1 capability levels, plus two C2 capability levels, plus two C3 capability levels, and one C4 capability levels assuming that for a level to function its prior level must function. Before switching parts we had two C1, and one C2 and one C3. After the switch is the more valuable more ready condition? We need some way to sum the values of capabilities when we have several aircraft and different combinations of capabilities possible. For example is four C3 better than two C3 and one C4. We might get the C4 if we used more money for part 8 and did not buy 2 part 6. Kaczynski et al. (2003) surveyed commanders and mechanics to see if there was a hierarchy of value and if it could be defined by a utility curve. If so then modeling could be done and optimized ordering and maintenance actions could be done with constraints on labor and money.

31.3.1 Modeling Example

Let us consider a basic structure. Referring to Figures 31.8 and 31.9, we consider a basic maintenance policy of purchase and transferring parts. A part set would be a rotor and an engine and gearbox which is a level one capability. For a helicopter with this part set a daytime humanitarian mission can be done. The variable y refers to capability at level j on aircraft k. In Figure 31.9 the incremental value (utility) is to C1: 0.4, to C2: 0.3, to C3: 0.2 and to C4: 0.1. This is modeled by the Greek letter delta in the mixed integer programming model below. If this curve is chosen the mission has a low probability of an attack requirement.

H1

7	0
5	6
3	4
1	2

H2

7	0
5	6
3	4
1	0

H3

0	8
0	0
0	0
1	2

FIGURE 31.7 Helicopter status example.

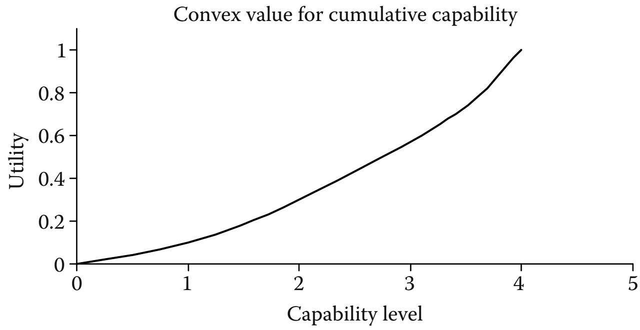


FIGURE 31.8 Example convex utility curve.

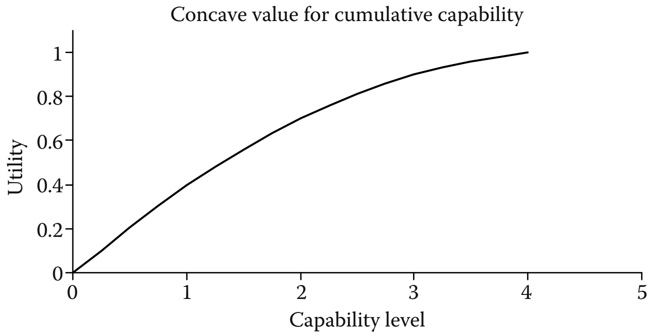


FIGURE 31.9 Example concave utility curve.

Variables:

$$y_{jk} = \begin{cases} 1 \\ 0 \end{cases}$$

Δ_j = value gained when parts set j is feasible on an aircraft k.

$x_{ijk} = \begin{cases} 1 \\ 0 \end{cases}$ if item i in parts set j is operable on aircraft k.

n_i = number of items in parts set i.
Q_i = on hand inventory for item i.
Objective function:

$$\text{Maximize} \sum_{j,k} y_{jk} * \Delta_j \tag{31.1}$$

Constraints:

$$x_{ijk} = 1 \tag{31.2}$$

$$y_{jk} \leq \left(\sum_i x_{ijk} \right) \Big/ (n_i) \tag{31.3}$$

$$y_{jk} \geq y_{j+1,k} \tag{31.4}$$

$$\sum_k x_{ijk} \leq Q_i \; \forall \; j \text{ and } i \text{ and for } i, j, k \text{ such that } x_{ijk} = 0 \tag{31.5}$$

Equation 31.1 sums all the value gained when a part set is feasible and complete. Equation 31.2 states a part designation is 1 when operating and 0 when nonoperable and *ijk* represents a part on level *i* the *j*th part on aircraft *k*. Equation 31.3 utilizes the fact that y is a 0 or 1 variable. If y is less than 1 it will be 0. This constraint assures y is not 1 unless all parts are operating. If there are three parts at a level and all are operating then $(1+1+1)/3=1$ and y for the next level is allowed to be 1. If there are two parts at a level then the y for the next is less than $(1+1+0)/3=2/3$ or 0. Equation 31.4 assures that a level is not feasible $(=1)$ when the level below is not feasible. If the y for the second level is 0, this forces the y for the third level to be 0. Equation 31.5 forces the replacement from the inventory to be less than or equal to the amount available.

Example

We have two helicopters. They have two parts at each level. H1: (1,0), (3,4), (5,6), (7,8); H2: (1,2), (3,0), (0,0), (0,0). In this model there is no cost constraint. If inventory for part 2 is zero, then the part in H2 will be transferred to H1 and a value of 1 is obtained as opposed to a value of 0.4 if the utility curve in Figure 31.9 is the value curve or 0.1 if the curve in Figure 31.8 is used. If we have an inventory of one for part 2, then the inventory part is used and the value is 1.1 if the curve in Figure 31.8 is used.

31.4 Testing Battery Life using Deterministic Spreadsheet Modeling

Modeling is often used to assist in describing and experimenting with real world systems. Models come in four basic types: Conceptual, physical, graphical and mathematical. Physical model are physical replications of the actual system (e.g. model cars, wind tunnels, etc.). Graphical models use drawings to represent the relationships between components of a system. Mathematical models are equations which represent the quantitative relationship between variables representing the system (West et al. 2008). Conceptual models are often tested and refined using mathematical models. There are numerous tools on can use to mathematically model deterministic conceptual systems. A tool readily available on most modern computer systems is a spreadsheet.

Spreadsheets can assist users in conducting basic examination of how systems work or could work, given fundamental data. An example of this is an experiment conducted by a team of West Point Cadets in 2005. Amongst other issues, the team studied the factors for power sources for remote sensors for a Disposable, Air-Droppable, Meteorological Tower Array (DAMTA) (Beissinger et al. 2005). Using data collected from real world tests (Table 31.1), and a basic understanding of electrical consumption, the team created a spreadsheet model of power consumption. The model was used to demonstrate the power consumption rate of the system over varying physical and environmental conditions.

With the data collected, the team calculated the current rate of power consumption based on system parameters and utilization procedures, the team discovered the current power sources were unable to meat the design requirements of 30 plus days. The Cadets used a spreadsheet model to modify parameters of the systems power sources and were able to identify the power capabilities and usage procedures required to provide a theoretical solution to the problem. The team calculated the likely power

TABLE 31.1 Example Power Consumption Results

Camera	Volts (V)	Amps (mA)	Amps (A)	Power (Watts)	Operating Time Est (Hours)
Camera Type A	11.87	23.62	0.0236	0.2804	84.93
Camera Type B	11.36	101.70	0.1017	1.1553	70.05
Camera Type C	11.43	91.70	0.0917	1.0481	68.55
Camera Type D					57.47

Source: Based on Beissinger, R., Horvat, P., Nelson, R., Schwab, J., and Powell, R. 2005. Disposable, air-droppable, meteorological tower array imagery integration. Unpublished report prepared for Army Research Laboratory. Department of Systems Engineering, United States Military Academy, West Point, NY.

source life expectancy for a single battery was sixteen and a half days if it ran the system at the minimal required time of 3 min per cycles and four cycles an hour (Table 31.2). This as about half the time specified for the system requirement of 30 days.

Searching for alternative solutions, the team's recommendation was to double the number of batteries which resulted in an average calculated operating time of 33.8 days of operation (Table 31.3). This met the systems design requirements and provided a 10% buffer.

Sensitivity analysis was conducted to see if other proposed components of the system would be viable based on likely power generation cells (Figure 31.10). Each factor in the design of the sensor systems had a global value (sum of which was 1.0) based on the opinions of subject matter experts. The subject matter experts believed the global value of the power source in the system design was 0.25. Figure 31.10 shows a current global value of 0.25 and a standard deviation of 10% (0.15 to 0.35), the Camera Type A has the best power consumption assessment. In fact, unless the global weight for power consumption falls below 0.10, the Camera Type A continues to outperform the other three alternatives.

This example illustrates how a spreadsheet model can be used as a viable tool for quick assessment of alternatives for a deterministic conceptual model of a system. From calculating the basic power consumption of a system to extrapolation the required power consumption and even sensitivity analysis of systems based on the global value placed on components of a system, the spreadsheet was able to create information from data for the decision makers.

The strengths of the spreadsheet for such assessments is its outstanding replication of actual system given real world data, the repeatability of the experimentation and the relative ease of verifying and validating the mathematical model as one can see the data and calculations. Weaknesses of using the spreadsheet to mathematically replicate the theoretical model is the limited response surface one can visualize and reasonable explore. Another weakness is the limited ability to replicate the randomness of the real world system across limited response area.

TABLE 31.2 Example Operating Time Calculations—DATMA Operating on one Battery

	Time (sec)	Time (Hours)	Volts (V)	Amps (A)	Power (Watts)
Low-Resolution Picture Cycle					
Transmiting	40.50	0.0113	11.34	0.43	4.8762
Idel	139.50	0.0388	11.64	0.37	4.3068
Total	180.00				
Total Amphours per cycle	0.0192				
Total Amphours per hour	0.0767				
Calculated Operating Time (Hours)	417.21				
High-Resolution Picture Cycle					
Transmiting	141.00	0.0392	11.25	0.44	4.95
Idel	39.00	0.0108	11.64	0.37	4.3068
Total	180.00				
Total Amphours per cycle	0.0212				
Total Amphours per hour	0.0850				
Calculated Operating Time (Hours)	376.62				
Amp Hours:	32				
Average Operating Time (Hours):	396.91				
Average Operating Time (Days):	16.54				

Source: Based on Beissinger, R., Horvat, P., Nelson, R., Schwab, J., and Powell, R. 2005. Disposable, air-droppable, meteorological tower array imagery integration. Unpublished report prepared for Army Research Laboratory. Department of Systems Engineering, United States Military Academy, West Point, NY.

Note: DATMA only funtioning during four three minute periods an hour with BW Camera.

TABLE 31.3 Operating Time Calculations—DATMA Operating on Two Battery

	Time (sec)	Time (Hours)	Volts (V)	Amps (A)	Power (Watts)
Low-Resolution Picture Cycle					
Transmiting	40.50	0.0113	11.34	0.43	4.8762
Idel	139.50	0.0388	11.64	0.37	4.3068
Total	180.00				
Total Amphours per cycle	0.0192				
Total Amphours per hour	0.0767				
Calculated Operating Time (Hours)	834.42				
High-Resolution Picture Cycle					
Transmiting	141.00	0.0392	11.25	0.44	4.95
Idel	39.00	0.0108	11.64	0.37	4.3068
Total	180.00				
Total Amphours per cycle	0.0212				
Total Amphours per hour	0.0850				
Calculated Operating Time (Hours)	753.24				
Amp Hours:	64				
Average Operating Time (Hours):	793.83				
Average Operating Time (Days):	33.08				

Source: Based on Beissinger, R., Horvat, P., Nelson, R., Schwab, J., and Powell, R. 2005. Disposable, air-droppable, meteorological tower array imagery integration. Unpublished report prepared for Army Research Laboratory. Department of Systems Engineering, United States Military Academy, West Point, NY.

Note: DATMA only funtioning during four three minute periods an hour with BW Camera.

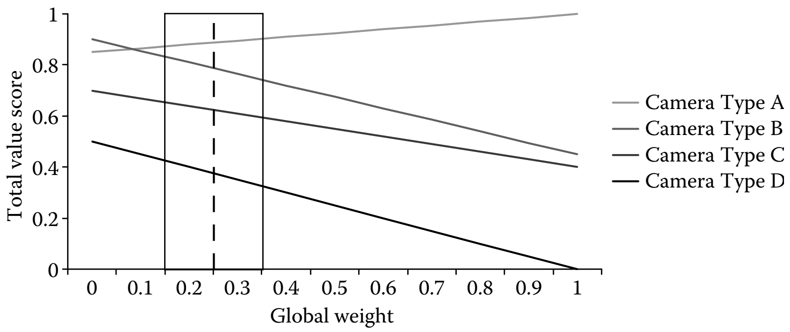


FIGURE 31.10 Example power consumption sensitivity analysis. (Based on Bedworth, D., Henderson, M.R., and Wolfe, P.M. 1991. *Computer Integrated Design and Manufacturing*. McGraw Hill, Inc., New York, NY.)

31.5 Modeling and Simulating Anti-Ballistic Missile Systems using Queuing Theory

As an organization of highly adaptive and innovative thinkers, the military is often constrained in its actions through doctrine and processes. This is especially true when placed in the context of highly lethal or strategic military action which is bond by a series of checks and balances designed to ensure appropriate use of these weapons systems. One of these highly structured weapons systems is Anti-Ballistic Missile (ABM) to defeat Intercontinental Ballistic Missiles (ICBMs). As such, the possibilities for modeling the processes to utilize these weapon systems are much easier then one might expect. True, it is a complex system with many variables, however, the manner in which these variables can or do interact is confined within the bounds of these processes.

RAMJETs are engines which provide a means of high speed flight. They use a relatively simple light weight propulsion system. The supersonic version of the RAMJET is known as the SCRAMJET (Benson 2006a, b).

The SCRAMJET, as well as the RAMJET, engine has no moving parts. The SCRAMJET (Figure 31.11) uses hydrogen gas uniting with oxygen in the air it extracts from the atmosphere to provide thrust. The engine operates best 20 miles and more above the surface of the earth. This results in a flight path different from missiles powered by more traditional technologies. Simulation provides a viable means to test the advantage(s) and disadvantage(s) these technologies present (Foote and Goerger 2005).

A theoretical model of the ABM engagement process is one of the first components required to model this system. In the model of the ABM command and control (C2) system developed by Hall et al. (2005), enemy missile(s) enter the system as entities which can be detected, handed to a C2 facility to be processed for potential engagement, the C2 element determines which of the ABM are viable for use in countering the attack (Figure 31.12). The C2 process passes the entity to one or more of three ABM weapons for potential engagement. Each phase of the process can result in the loss of the opportunity to engage the threat missile system(s).

The process flow describes the theoretical model for a structured engagement of the ICBMs. To exercise the model and add fidelity to this model, a series of data are required. Some possible parameters include:

- Specifics of launch threat
    - Flight characteristics
    - Possible launch location
    - Potential targets
- Target specifications
    - Location from time sensitive target
    - Range of target
    - Time available to strike target
- Command and control
    - Time to process data
    - Time to make decisions
    - Time to communicate decisions
- SCRAMJET or RAMJET specifications
    - Time to launch
    - Shock effects and air turbulence effects
    - System accuracy
    - System speed
    - System acceleration
    - System location

FIGURE 31.11 SCRAMJET. (From Benson, T. 2006. SCRAMJET Propulsion, http://www.grc.nasa.gov/WWW/K-12/airplane/scramjet.html, Glenn Research Center. With permission.)

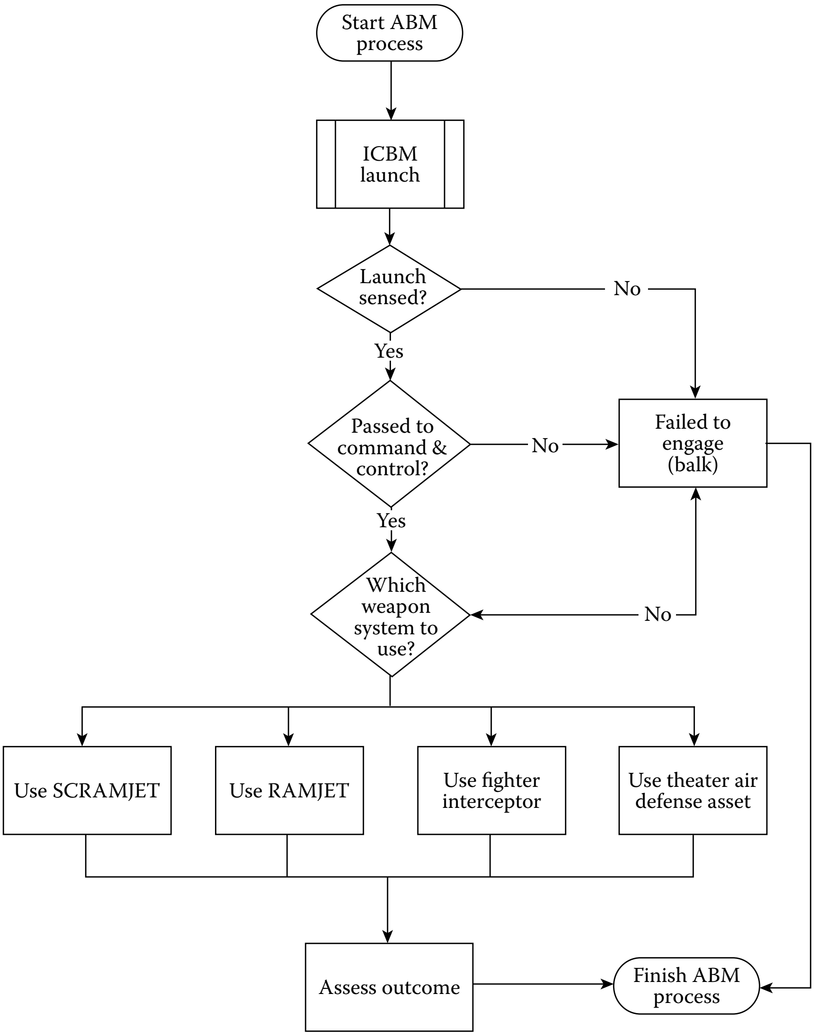


FIGURE 31.12 Anti-ballistic missile (ABM) engagement process flow chart. (Based on Hall, J., Mayfield, C., Toll, R., and Yu, D. 2005. Impact of SCRAMJET technology in supporting the Army's mission to defend CONUC against cruise missiles and TBM. Proceedings of the 2005 Systems and Information Engineering Design Symposium, Virginia, Charlottesville, VA, April 29, 45–54.)

    - System range
    - System acquisition of target
    - Time available to strike threat
    - Probability of hit/kill (PH/PK)
- Multiple launch capability
- Contingent effects of action
- Assumed weaponization structure specifics for new system

The objective of the Hall et al. study was to determine if SCRAMJET and RAMJET technology increased the probability of ABM success in a variety of scenarios. To accomplish this, the model had to produce results which would assist in quantifying the value of using or having a multitude of options to include SRAMJET and RAMJET technologies. If the model indicated the probability of a layered defense significantly increased the performance of the ABM process then it could infer that one could quantify the value of these new weapon systems. If the probability of destroying a time sensitive target significantly increased then one would be able to see the value of process and technology in the system.

31.5.1 Simulation Methodologies

Simulations are models run over time. They are often used to provide insight to the possible effects of changes in time or variables (West et al. 2008). Modeling a system is effective in providing insight or options to the decisions maker. However, run the model or a collection of models over time is required to more clearly identify the interactions and possible effects of a complex system.

One can simulate and thus potentially show the impact of a system using one of many mathematical modeling tools. For instance, one seeks to represent the flight path of an ICBM, they can use a parabolic based model. However, as a semblance of the real world, at some level of fidelity, the model no longer accurately describe the actual trajectories of an ICBM. This can be seen as one models and simulates the ICBM's path of flight. Habitually, ICBMs follow an elliptical trajectory with one of the two foci at the earth's center. This trajectory permits the ICBM to obtain the greatest velocity and range possible. However, a model can accurately describe the ICBM's trajectory during the critical moments of its flight, launch and terminal approach, while using a less computationally complex and thus possibly less accurate flight model for the time the missile is flying through space. History has shown that missiles are most vulnerable during launch and just prior to impact. Hence, the parabolic models produce a viable trajectory representation. Since an ICBM is most vulnerable to hypersonic interceptors at these two moments, this is where one would seek to provide model fidelity (Hall et al. 2003).

The stochastic nature of a system is often replicated in a spreadsheet model using random values for some of the variables over time to simulate system activity. In cases where a limited number of random variable (4–5) with a limited number of ranges in value (4–5 each) one can possible use a spread sheet to model the effects and enumerate the exhaustive set of parameter combinations. In a case where one has five variables with a range of five possible values each, this is a spread sheet that accounts for 5! or 120 options.

31.5.2 Stochastic Spreadsheet Assessment

One method of simulating a system such as the ABM is the use of a spreadsheet queuing model. Based on the ICBM scenario, Hall et al. created a spreadsheet simulation of the SCRAMJET and ABM process to help identify possible the probable outcomes of an engagement. In Table 31.1 we see how the Cadets took basic physics algorithms/models, a simple variable arrival rate, variable service time, single server queuing model (using Kendall Notation–M/M/1)* (Womack et al. 2003) and random values for distances and service times to create a list of possible service times to intercept ICBMs using weapons that used SCRAMJET technology. This use of random numbers produced stochastic results to help identify the likely boundaries of performance for the system. The examination of four randomly generated attacks over time (bottom four lines of Table 31.4), simulates the ABM system and prides one possible set of inputs and outcomes.

Using the basic formula for the average time in queue and waiting time including service (Equations 31.1 and 31.2, respectively), the team was able to illustrate and determine the average service time required for a ABM system using SCRAMJETs to deploy and destroy an ICBM. This provided bases for assessing when and where the SCRAMJET would likely intercept an ICBM and thus whether this provided enough time and space to employ multiple systems before the ICBM could breach friendly territory.

$$W_q = \frac{L_q}{\lambda} = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)} \tag{31.1}$$

* M =(Markovian") Poisson or exponential distribution.

TABLE 31.4 Example Excel Based Queuing Model

Pre Fire Components:	Time(sec)	Mean(sec)	Standard Deviation(sec)		S =	841.36	km			
Identify:	2.85	3	0.5							
Orient:	3.26	5	1		Mach 1 =	331.4599988	m/sec			
Report:	15.34	15	5		Mach	4	=	1325.84	m/s	
Reply from Higher:	23.68	25	5							
Fire:	2.36	4	1		Time of Level Flight	634.6	sec	Time to Intercept	317.3	sec
Identify to Fire Time:	44.23		3			10.58	min		5.29	min
						0.176	hours		0.088	hours
Max Detection Range:	900	km								
Dead Waste:	58.64	km			Mass:	2000.00	kg			
Actual Response Range:	841.36	km			Velocity:	1325.84	m/sec			
					Kinetic Energy:	1757851693	J			
Time to interception:	317.3	sec			Bowl Size:					
Service Time:	361.52	sec								
	6.03	min								
Input Parameters:		Inter-arrival Time (IA) (Exponential)	λ =	0.100	(incoming missles/sec	$1/\lambda$ =	10.000	(sec)		
		Station 1 Service Time (Normal)	mean =	sec	(sec)	σ =	100.000	(sec)		
		Avg IA Time			Wq1		Average Svc Time (Station 1)			
	.=AVERAGE(C27:C30)	8.61	sec		52.17	sec	361.33	sec		
1	2	3	4	5	6	7	8	9	10	
Incoming Missile #	Randome # for Interarrivals	Inter-Arrival Time (sec)	Arrival Time (sec)	Start Service Time (sec)	Wait time (Battery 1) (sec)	Random # for Service Times	Service Time (sec)	Missile Destroyed Time (sec)	Identify to Fire Time (sec)	
1	0.40	5.18	5.18	5.18	0.00	0.61	388.33	393.51	44.49	
2	0.87	20.28	25.46	44.49	19.04	0.66	401.56	446.05	45.68	
3	0.23	2.59	28.05	45.68	17.63	0.69	412.09	457.77	47.14	
4	0.45	6.00	34.05	47.14	13.09	0.60	387.12	434.26	46.98	

Source: From DAZ Productions, 3D Models, 3D Content, and 3D Software, http://www.daz3d.com/program/bryce/55index.php.

$$W = W_q + \frac{1}{\mu} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu - \lambda} \quad (31.2)$$

where L = expected number of customers in the system; L_q = expected number of customers in the queue; W = expected waiting time (including service); W_q = expected waiting time in the queue; ρ = system or resource utilization; P_n = probability of having n customers in the system; s = number of servers; λ = mean arrival rate; $1/\lambda$ = mean inter-arrival time; $1/\mu$ = mean service rate; μ = mean service time.

Follow on work goes on to calculate the probability of intercepting ICBMS based on specific locations in the world. Hall et al. also created a spreadsheet model to calculate the likelihood that a SCRAMJET would intercept an ICBM based on general characteristics of both the site being defended and the ICBM.

This example illustrates again that with sound fundamental spreadsheet skills and viable algorithms to represent system behaviors, a spreadsheet can be used to model basic effects of a system. We see the strengths of the spreadsheet to quickly test a stochastic mathematical model with more fidelity then other traditional methods. However, the spreadsheet is still limited in the number of variables, the ability to make runs and replications and a lack of visualization to help illustrate the dynamic nature of a system. In general, spreadsheet are an excellent means of gaining confidence in the fundamental conceptual model before speeding more extensive resources to collect more data on the system and enhance the fidelity of the model.

31.5.3 Stochastic Simulation Software

Another, more dynamic method of modeling a system such as the ABM is the use of a commercial queuing simulation. The use of this simulation helped provide possible optimal solutions to a multidimensional response surface which maybe to vast for a spreadsheet simulation to execute efficiently. Simulation packages such as this can often provide a systematic and potentially exhaustive search of the response surface.

To test this premise and demonstrate its ability to proved insights to complex military weapons systems, ABM simulation was developed in ProModel (Figure 31.13) based on a slightly more complex ABM engagement process based on the Hall et al. (Figure 31.14). In this simulation, targets enter the system as entities which can be detected, passed to a command and control facility to be processed for potential engagement, and then passed to one of three ABM weapons for potential engagement. Each phase of the process can result in the loss of the target missile system. Only the weapons systems can engage and destroy the target missile system.

Utilizing this theoretical model and a commercial of the shelf (COTS) simulation package, a three layer weapons defense against ballistic missiles was investigated through a more robust range of alternatives then was practical with the spreadsheet model. The results of the simulation package were used to determine the effectiveness of potential engagement protocols. Due to limited real world data of direct engagements using ABM and the developmental nature of the SCRAMJET weapon system, most of the modeled probabilities are assumed and varied in accordance with projected surrogation with existing ABM systems.

Employing a queuing simulation tool, one is able to investigate potential protocol priorities based on available weapon systems, time to target, probability of kill, and time to impact. One of the results of such an investigation is a set of protocols able to minimize the time to make critical employment decisions which in turn could increase the likelihood of a successful engagement of ICBMs.

With simulation packages such ProModel, analysts can examine the system with greater fidelity then spreadsheet simulations, use more complex variables, execute more runs and replications, easily replicated experiments. Textural output files provide a means to test repeatability and validate simulation processes and results. The simulation also provides visualization to assist in verifying the models adherence to the conceptual model and rapidly illustrate to decision makers the complexities of the system in a manner traditionally easier to assimilate. However, such simulations are routinely more time consuming to produce then a spreadsheet simulation and require vast amounts of data to create viable distribution curves to replicate reality for the higher fidelity stochastic model.

Analyzing complex weapons systems does not require the use of system specific or combat simulations to provide valid insights and recommendations. The use of products such as spreadsheets or COTS

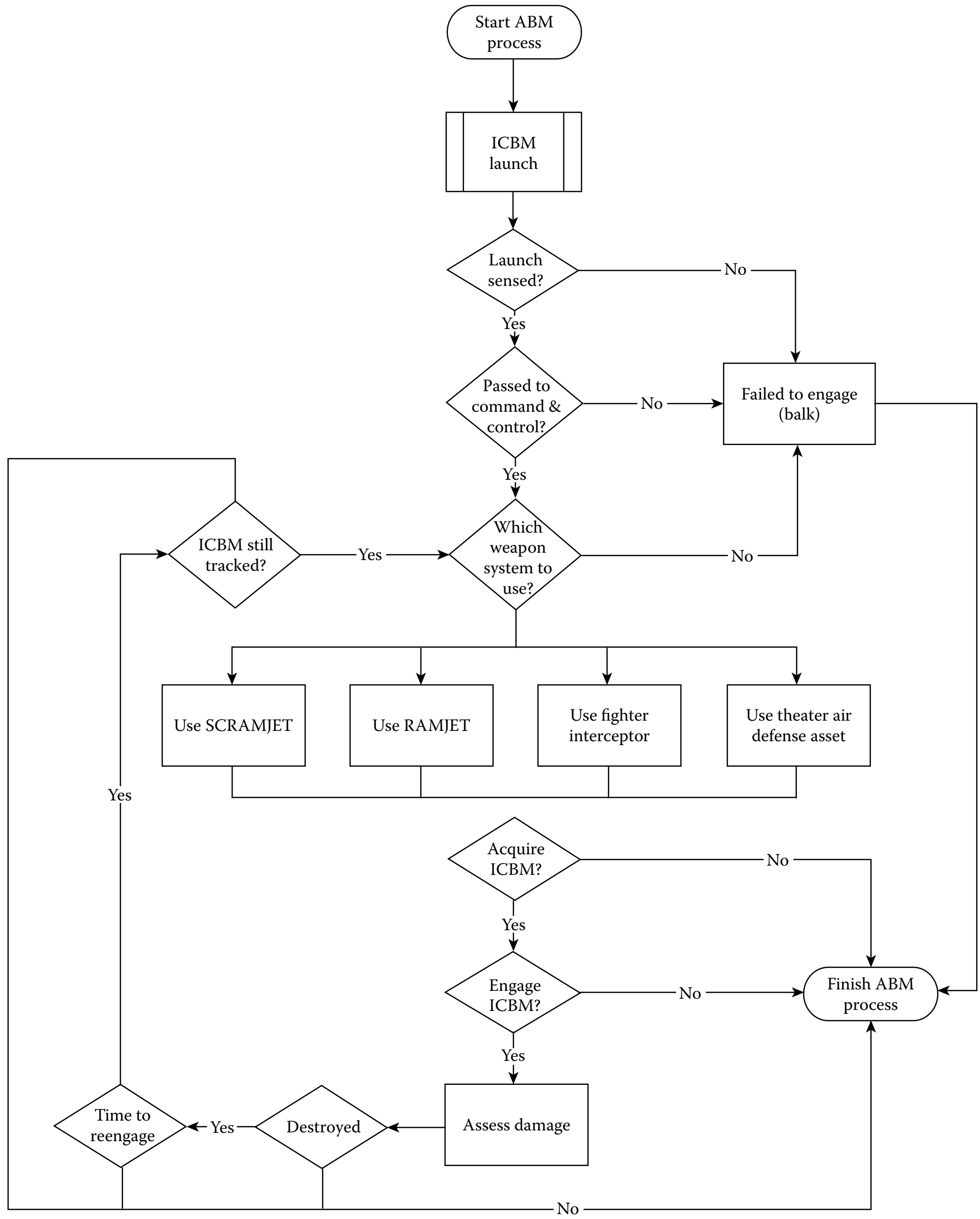


FIGURE 31.13 Anti-ballistic missile (ABM) engagement process flow chart II. (Based on Hall, J., Mayfield, C., Toll, R., and Yu, D. 2005. Impact of SCRAMJET technology in supporting the Army's mission to defend CONUC against cruise missiles and TBM. Proceedings of the 2005 Systems and Information Engineering Design Symposium, Virginia, Charlottesville, VA, April 29, 45–54.)

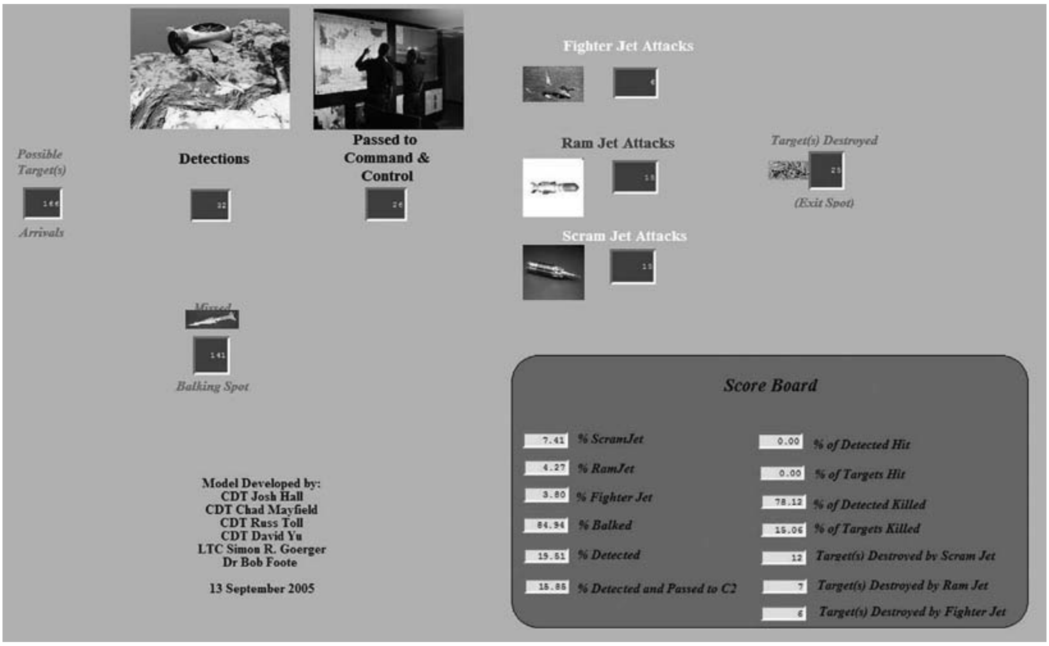


FIGURE 31.14 ProModel anti-ballistic missile simulation example. (From Foote, B.L. and Goerger, S.R. 2005. Design considerations for simulating ABM systems. Presentation, Huntsville Simulation Conference, October 26–27, Huntsville, AL. With permission.)

simulations packages can to provide decision makers with basic insights for systems such as ABM systems. Using spreadsheets or COTS simulations packages requires less time to code viable scenarios and can provide fundamental insights or help focus efforts in the analysis of more complex interactions which can only be investigated using system specific or combat simulations.

31.6 Testing Power Management using Agent Based Modeling

Previously we showed how spreadsheets can assist in determining if a battery type can meet current and required power source requirements. We have seen how seen how a commercial simulation package can assist in determine appropriate procedures and policies for the implementation of a developmental weapon system. This section demonstrates how an agent passed simulation can be used to identify key sub systems utilization to help determine power management policies to reduce power consumption.

As our soldier systems become more and more technically advanced, relative to mechanical advanced, power requirements become increasingly significant to ensure effective operations. Many of these power requirements are fulfilled though the relatively insufficient power storage devices such as batteries. One developing weapon system highly reliant on battery power is the Land Warrior (LW) which is a Soldier Tactical Mission System (STMS) (Womack et al. 2003). The LW is a series of weapons, communications, and protective systems transported and operated by the dismounted soldier. Thus, the power storage devices that operate the electronics in these subsystems are also carried by the soldier. With the physical limitations of the human being and the relatively incremental development of batteries, researchers are looking at power management as a means of reducing power consumption and thus extending the usability of systems without increasing the number of power sources carried by the LW.

An essential element in developing a power management architecture is identify the key sub systems or component and under what qualitative conditions they are required to achieve mission success. Once this is accomplished, analysts can conduct studies to provide decision makers with viable recommendations for appropriate tactics, techniques and procedures for the utilization of theses key components. In turn, this will reduce power consumption and extend the usability of the electronic component with out requiring addition power storage devices to be carried and used by the LW.

Determining the utility of the LW electronic components was the focus of a study conducted by three West Point Cadets in the spring of 2003: CDT Scott Womack, CDT Dan Mcconnell, and CDT Alaina Reese (Womack et al. 2003). The Cadets used complexity theory and Agent Based combat simulation to examine the STMS. There work was used to help develop a conceptual framework for the use of more traditional combat simulations to analyze these components over a wider set of scenarios, with greater system fidelity and in conjunction with other combat systems. The Operational Requirements Document for LW identified six unique scenarios for which STMS had to operate. The cadets focused one of these scenarios, assault a simple objective, to examine eight essential LW systems: Thermal Weapons Sights, Multi-Functional Laser, Daylight Video Sight, Night Optics, Dead Reckoning Module and Global Positioning System, Digital Map Display, Radio, and Individual Combat ID for the Dismounted Soldier (Bailey et al. 2003).

The cadets developed the offensive study scenario in MANA (Womack et al. 2003), and agent based simulation developed by the New Zealand Defense Ministry and utilized by the US Marine Corps in Project Albert. Figure 31.15 illustrates base case scenario in MANA. Although MANA did not provide the desired level of fidelity for all the systems, it provides the best level of fidelity then the other agent based simulations at the time.

The experimental design for the study contained seven parameters: communications delays, number of living enemy entities, ease of movement by friendly force, sensor range, movement speeds for friendly soldiers, clustering of soldiers, and combat Force ratio required to maintain offensive action (Womack et al. 2003). These seven parameters allowed for the assessment of required LW systems through inference of the capabilities that certain systems provide. Fore example, the effectiveness of the digital map and radios allows for virtual clustering of the forces versus a requirement for physical clustering based on line-of-site.

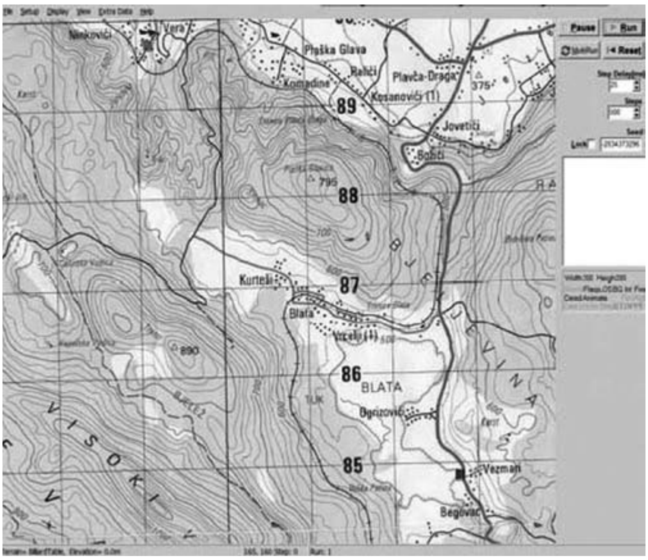


FIGURE 31.15 Base case scenario in MANA. (From Deak, M.Y. and S.S. Hoon. 2003. Analysis for load-balancing techniques in intelligent network systems. *Simulation*, 79(1). With permission.)

Analyzing data generated from the MANA 34, 560 different excursions, Womack et al. identified sensor range was the played the greatest impact on mission success. Interactions between sensor range and movement speed, number of living enemy, clustering, and force ratio (Womack et al. 2003).

The results of the study demonstrated that agent based models can be used to explore a large experimental landscape identifying key factors to be investigated using more complex combat simulations (Womack et al. 2003). This allows for a more comprehensive exploration of alternatives and provides insight to key components for development of process. In this incident searching for key systems required to ensure mission success, when the each system is required, when each system should be used, and conversely where each system could be placed in a power saving mode to preserve power for future needs (Sanders 2003).

References

Bailey, M., Sanders, D.M., and Arnas, A.O. 2003. Meeting military energy demands. Proceedings of the 2003 American Society of Mechanical Engineers Thermal Engineering Joint Conference, Orlando, FL.

Bedworth, D., Henderson, M.R., and Wolfe, P.M. 1991. *Computer Integrated Design and Manufacturing*. McGraw Hill, Columbus, OH.

Beissinger, R., Horvat, P., Nelson, R., Schwab, J., and Powell, R. 2005. Disposable, air-droppable, meteorological tower array imagery integration. Unpublished report prepared for Army Research Laboratory. Department of Systems Engineering, United States Military Academy, West Point, NY.

Benson, T. 2006a. RAMJET/SCRAMJET Thrusts. http://www.grc.nasa.gov/WWW/K-12/airplane/ramth.html, Glenn Research Center (Accessed 18 September 2007).

Benson, T. 2006b. SCRAMJET Propulsion. http://www.grc.nasa.gov/WWW/K-12/airplane/scramjet.html, Glenn Research Center (Accessed 18 September 2007).

DAZ Productions, 3D Models, 3D Content, and 3D Software. http://www.daz3d.com/program/bryce/55index.php (Accessed 13 September 2005).

Deak, M.Y. and Hoon, S.S. 2003. Analysis for load-balancing techniques in intelligent network systems. *Simulation*, 79(1), 35–42.

Foote, B.L. 1972. The sensitivity to number of periods of a multi-period no set up stochastic inventory policies. *Industrial Engineering*, March, 65–68.

Foote, B.L. and Goerger, S.R. 2005. Design considerations for simulating ABM systems. Presentation, Huntsville Simulation Conference, October 26–27, Huntsville, AL.

Hall, J., Mayfield, C., Toll, R., and Yu, D. 2005. Impact of SCRAMJET technology in supporting the Army's mission to defend CONUC against cruise missiles and TBM. Proceedings of the 2005 Systems and Information Engineering Design Symposium, Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, April 29, 45–54.

Irani, S.A. 1999. Handbook of Cellular Manufacturing Systems. Wiley-Interscience, Norcross, GA.

Kaczynski, W.H., Foote, B.L., Pohl, E.A. 2003. , "A Utility Based Optimal Metric Coordinating Mission Capability and Supply Level," Proceedings of the Institute of Industrial Engineers Annual Research Conference, Houston, Texas, 16 May 2003.

Kebriae, N. 1981. A study in sensitivity analysis of a dynamic random lead-time inventory model. Masters Thesis, University of Oklahoma, School of Industrial Engineering, University of Oklahoma, Norman, OK.

Liberatore, M.J. 1977. Planning horizons for a stochastic lead-time inventory model. *Operations Research*, 25(6), 977–988.

McCauley-Bell, P.R. and Crumpton-Young, L.L. 2006. Introduction to applications of fuzzy set theory in industrial engineering. In Badiru, A.B., editor, *Handbook of Industrial and Systems Engineering***.** Taylor and Francis, Boca Rotan, FL, 25-1–25-18.

Sanders, D.M. 2003. *Land Warrior Power Management*. Defense Technical Information Center (DTIC) number ADA414902. Department of Systems Engineering, United States Military Academy, West Point, NY.

West, P.D., Kobza, J.E., and Goerger, S.R. 2008. Systems modeling and analysis In: Parnell, G.S., Driscoll, P.J. and Henderson, D.L., editors, *Systems Decision Making in Systems Engineering and Management*. John Wiley & Sons, Inc., Hoboken, NJ.

Womack, S.F., McConnell, D., Reese, A., and Sanders, D.M. 2003. Power management for the Land Warrior system. In *Systems and Information Engineering Design Symposium, 24–25 April*. IEEE, 83–88.

Zadeh, L. 1975. Fuzzy Sets and Applications to Cognitive and Decision Making Processes**.** Academic Press, NY.

32

Countering Forgetting Through Training and Deployment

Mohamad Y. Jaber
Ryerson University

Hemant V. Kher
University of Delaware

Darwin J. Davis
University of Delaware

32.1 Introduction

The benefits of a flexible workforce, whereby workers are cross-trained to operate machines in functionally different departments, have been well documented in practice (Wisner and Siferd 1995) and in the DRC literature. Assigning workers to different departments is useful in alleviating the detrimental effects of bottlenecks resulting from machine breakdowns, product type changes, or external demand changes. A flexible workforce can reduce work-in-process inventory levels and lead times, and improve customer service performance (Park and Bobrowski 1989), but acquiring worker flexibility is costly. In the process of training workers in different departments, shops typically incur productivity losses due to learning effects. Park and Bobrowski (1989) provide specific examples of training related costs—Chrysler Corporation used over 900,000 hours before launching the new Jeep, and GM provided each worker over 80 hours of training before starting new product lines.

Shops also incur on-going costs to "maintain" worker flexibility due to forgetting which results from interruptions in the learning process. Wisner and Pearson (1993) surveyed managers in six different firms. Five of these managers attested to the presence of relearning effects associated with worker transfers. Three managers used period retraining policies with the objective of reducing relearning effects. A recent survey of make-to-order machine shops based in the US indicates that managers are concerned with losses associated with workers relearning what they have forgotten as they transfer among different processes in the shop (Wisner and Siferd 1995; Wisner 1996). These surveys list strategies used by these firms to reduce forgetting.

The importance of training and deployment policies that address the worker efficiency versus flexibility tradeoff is amply evident from the surveys of Wisner and Siferd (1995) and Wisner (1996). McCreery and Krajewski (1999) also emphasize this by stating: "*in our discussions with a number of plant managers, there is a great deal of concern on how best to manage the problem of developing efficiency through worker specialization versus developing worker flexibility*". In this chapter, we study worker cross-training and deployment using the learn-forget curve model (LFCM) developed by Jaber and Bonney (1996). Related work by Wisner and Pearson (1993), Ash and Smith-Daniels (1999) and McCreery and Krajewski (1999) examines worker deployment issues in the presence of learning and forgetting. Also, Kher et al. (1999) and Kher (2000) studied worker cross-training and deployment in the presence of learning and forgetting. While these studies acknowledge that forgetting significantly impacts shop performance, they use learning–forgetting models that do not fully capture the factors that influence forgetting in industrial settings. As shown later, in comparison to the models used in the above studies, LFCM more completely models characteristics associated with forgetting in industrial settings.

The remainder of this chapter is organized as follows. We review characteristics that influence forgetting as identified in previous literature, and evaluate the degree to which existing learning–forgetting models are consistent with these characteristics. We simulate worker learning and forgetting with LFCM to investigate training and transfer policies. Given that prior work by Kher (2000) and Kher et al. (1999) examines related issues using the variable regression variable forgetting (VRVF) model of Carlson and Rowe (1976), we contrast the LFCM results with those of the VRVF model. Finally, we consider an extension of LFCM to include the influence of task similarity on forgetting.

32.2 Literature Review

The following characteristics of forgetting have been identified in laboratory and empirical studies:

(1) *The amount of experience gained before interruption occurs in the learning process influences the level of forgetting*: Studies by Bailey (1989), Globerson et al. (1989), and Shtub et al. (1993) among others have established that in general, less experienced workers will likely forget more, whereas workers with greater experience will forget less.
(2) *The length of the interruption interval influences the level of forgetting*: Studies by Bailey (1989), Globerson et al. (1989), Shtub et al. (1993) and Dar-El et al. (1995a,b) are in agreement that for a given level of experience, a worker experiencing a longer break will forget more than one experiencing a shorter break.
(3) *Relearning rate is the same as the original learning rate*: Researchers have debated this issue. Results in Globerson et al. (1989), Sparks and Yearout (1990) and Shtub et al. (1993) support this finding. Bailey and McIntyre (1997) also find partial support. However, laboratory studies by Bailey (1989) and Hewitt et al. (1992) seem to indicate a lack of correlation between the original learning rate and the relearning rate.
(4) *The power-based model is appropriate for capturing forgetting effects*: Globerson et al. (1989) and Shtub et al. (1993) confirm this finding.
(5) *Learning and forgetting are mirror images of each other*: Globerson et al. (1989) support this behavior of forgetting in their study.
(6) *The level of forgetting depends upon the rate at which a worker learns*: Two recent studies (Nembhard 2000; Nembhard and Uzumeri 2000), based in an industrial setting with several learning–forgetting episodes for hundreds of workers, established that workers who learn rapidly also tend to forget rapidly.
(7) *The nature of the task being performed influences the amount of forgetting*: Studies by Sparks and Yearout (1990), Hewitt et al. (1992), Dar-El et al. (1995a,b), Arzi and Shtub (1997) and Nembhard (2000) attest to this finding. Results appear to be somewhat mixed in other studies. For simple tasks, forgetting appears to be related to the level of prior experience and the length of the

interruption interval. For complex tasks, the influence of break length is less clear. In some cases, breaks impact forgetting as expected, however, for certain cognitive tasks such as problem solving, or algorithm development (Arzi and Shtub 1997), performance following a break shows improvement rather than deterioration.

Models that fail to incorporate these important characteristics of forgetting have the potential to generate misleading results. For example, if a model does not account for prior experience, then for a given length of interruption, a worker with 10 years of experience would experience the same amount of forgetting as a worker with only 10 days of experience. Similarly, a model that ignores the length of interruption interval would suggest that a worker interrupted for 1 year would experience the same amount of forgetting as a worker interrupted for 1 day. Misleading results from a model could lead to the development of ineffective training policies. Specifically, with inadequate training, worker reassignment may result in substantial relearning. Conversely, excessive training requirements prevent managers from effectively utilizing worker flexibility.

We now compare existing learning and forgetting models against the characteristics #1–#7 noted above. Nembhard (2000) empirically validated a power based, four parameter learning–forgetting model which considers interruption duration and the recency of experience in calculating forgetting. However, this model ignores the amount of experience prior to the interruption (characteristic #1), does not model learning and forgetting as mirror images (characteristic #5), nor does it correlate the workers' forgetting rates with their learning rates or task complexity (characteristics #6–#7).

The VRVF, the variable regression invariable forgetting (VRIF) model of Elmaghraby (1990), and LFCM all satisfy characteristics #1–#4. These models account for the amount of prior learning and the interruption duration, they use a relearning rate equal to the original learning rate and they use power-based equations. However, they differ in calculating the extent of forgetting. If forgetting is in fact a mirror image of learning, as stated by Globerson et al. (1989) (characteristic #5), then, the forgetting function must satisfy two properties. First, when total forgetting occurs, the time required to process the next unit should equal the time needed to process the very first unit without any prior experience. Second, the performance time on the learning curve must equal that on the forgetting curve at the point of the interruption. Whereas LFCM is consistent with both the properties, VRIF is only consistent with the first and VRVF is only consistent with the second. Although VRVF and VRIF incorporate most all of the characteristics of forgetting, we note that VRIF has mathematical limitations that make it unsuitable for studying worker training (see Appendix at the end of this chapter).

A notable difference between these models is that VRVF and VRIF use a fixed, externally specified forgetting rate. In contrast, the LFCM calculates forgetting rate for each instance of forgetting, and the forgetting rate is expressed as a function of the learning rate. This allows the LFCM to capture the empirically observed phenomenon that workers who learn rapidly tend to forget rapidly (Nembhard and Uzumeri 2000), satisfying characteristic #6. Similarly, given that learning rates for complex and simple tasks vary (Dar-El et al. 1995a,b), the LFCM captures forgetting for these two task types differently, thereby satisfying #7.

While several studies have examined worker cross training and/or deployment policies, the models used in these studies fail to incorporate important characteristics of forgetting. For example, the model used by Wisner and Pearson (1993) does not satisfy characteristics #1 and #2. Ash and Smith-Daniels (1999) use a model that satisfies #2 but not #1. The model in McCreery and Krajewski (1999) satisfies #1 but not #4. For low levels of experience their model will overestimate the amount remembered, and therefore underestimate forgetting. Finally, Kher et al. (1999) and Kher (2000) use the VRVF model whose limitations were discussed above.

In summary, many factors are associated with the forgetting phenomenon. While other models exhibit some of these characteristics, the LFCM embodies the majority. Furthermore, in an evaluation of LFCM, Jaber and Bonney (1997) predicted the performance times from the laboratory experiment of Globerson et al. (1989) with less than 1% error.

Because of its superior adherence to these characteristics we wish to study training and deployment policies using LFCM. Specifically, in studying training and deployment we would like to know:

(A) Will providing a worker upfront training reduce forgetting?
(B) How does the frequency of worker transfers relate to forgetting?
(C) Do the answers to above questions A and B change as the number of tasks for which a worker is trained increases or for different forgetting rates?

These questions motivate our experimental design that is given in the next section. Because, similar work has been previously been done by Kher et al. (1999) using VRVF, we also include VRVF in our study to determine if there are notable differences in the results of these models. In general, we follow the experimental design of Kher et al. (1999).

32.3 Experimental Design

We simulate an environment in which a worker is learning to process jobs on different machines. With increasing experience, performance times decrease in accordance with the plateau log-linear learning model (similar to that proposed by Wright 1936):

$$T_n = \begin{cases} T_1 n^{-b}, & \text{if} \quad n \le n_s \\ T_s & \text{, otherwise} \end{cases} \tag{32.1}$$

where T_n is the time to perform the nth repetition (job), T_1 is the time to perform the first repetition (assumed to be four times the standard processing time), b is the learning slope ($0 \le b \le 1$), n_s is the number of units required to be performed to reach standard time, and T_s is the standard time. We use a deterministic simulation in which the shop has a constant availability of work.

We use the following four experimental factors: (1) the number of tasks for which the worker is being trained; (2) transfer policy that controls worker transfers between tasks; (3) the rate at which a worker forgets a task as a result of the interruption in task learning; and (4) the extent of upfront training that is provided to the worker. Details on the implementation of these factors are given below.

The number of tasks a worker learns in the shop is modeled at two levels. At the first level the worker is trained to process jobs in two departments, while at the second level the worker is trained to process jobs in three departments in the shop. A fixed batch size controls worker transfers in the shop. For example, in the case where a worker is learning two tasks (FLEX=2), she/he processes a fixed number of jobs at the first machine, and then proceeds to the second machine to process the same number of jobs. Batch sizes of 10, 25 and 250 jobs are used to reflect the trade off between frequent transfers (when using batch sizes of 10 or 25) and greater efficiency (when using a batch size of 250). A cycle is completed when a worker visits machines 1 and 2 when two tasks are being learned or machines 1, 2 and 3 when learning three tasks.

Forgetting rates are modeled at three different levels in our analysis ranging from 85% (high forgetting) to 90% (medium forgetting) to 95% (low forgetting). We refer to upfront training as the flexibility acquisition policy (FAP), and study it at three levels. In the base case (FAP-0), no upfront training is provided. Thus, worker transfers between machines in each cycle are controlled by the batch size under FAP-0. Under FAP-1 and FAP-2 policies, the worker must process 370 and 740 jobs in the first cycle, respectively. Processing 370 jobs in the first cycle results in the worker reaching standard processing time before being transferred from a department. Processing 740 jobs in the first cycle, the worker reaches standard processing time (at 370) and then continues along the plateau of the learning curve for another 370 jobs. Under the FAP-1 and FAP-2 policies, worker transfers between machines are controlled by the batch size from the second cycle onwards. After the worker has completed approximately 3000 jobs in each department, we calculate the average processing time over all jobs processed.

Worker's learning and forgetting effects are captured using the LFCM and VRVF models. We note that while VRVF uses a fixed forgetting rate in all the cycles, the forgetting rate for the LFCM is at the specified level only in the first cycle. For subsequent cycles, a new forgetting rate must be calculated for the LFCM. We will now elaborate on how forgetting effects are captured using the LFCM and VRVF.

32.3.1 Learning–Forgetting Models

The commonly used power-form forgetting curve is given by:

$$\hat{T}_x = \begin{cases} \hat{T}_1 x^f, & \text{if } d < D \\ T_1, & \text{otherwise} \end{cases} \tag{32.2}$$

where $\hat{T}_x$ is the time for the xth repetition (job) of lost experience of the forget curve, x is the number of potential repetitions (jobs) that would have been performed if interruption did not occur, $\hat{T}_1$ is the intercept of the forget curve, f is the forgetting slope, d is the length of the interruption period in any production cycle, and D is the time to which total forgetting is assumed to occur if the process was interrupted for a period of length d, where $d \leq D$.

Consider a production situation where interruption occurs immediately after processing the nth job. If the length of the interruption period is sufficiently long, then some of the experience gained in processing $\sum_{j=1}^{i-1} n_j$ jobs in the $i-1$ previous cycles is lost. As a result, the processing time for the first job in cycle $i+1$ will revert to a value higher than that at the point of interruption in cycle i.

32.3.2 Variable Regression Variable Forgetting (VRVF) Model

Carlson and Rowe (1976) assumed that the intercept of the forget curve, $\hat{T}_1$, varies in each cycle based on the value of v_i, while the forgetting slope remains constant, i.e. $\hat{T}_1 \neq \hat{T}_2 \neq \hat{T}_3 \neq \ldots \neq \hat{T}_i$ and $f_i = f_1 = f$ for $i = 1, 2, 3, \ldots$

Consider an intermittent production situation where interruption occurs after v_i repetitions (jobs) have been performed in cycle i, where $i = 1, 2, 3$, etc. The value of v_i represents the sum of the number of repetitions remembered at beginning of u_i, and in n_i, for cycle i, i.e. $v_i = u_i + n_i$. It is at this point that the forgetting function is defined by equating the time required to produce the v_ith repetition on the learning curve to the time required on the forgetting curve. This is done by equating Equations 32.1 and 32.2, and solving for $\hat{T}_1$ which gives:

$$\hat{T}_{1i} = \begin{cases} T_1 v_i^{-(b+f)}, & \text{if } v_i < n_s \\ T_s v_i^{-f}, & \text{otherwise} \end{cases} \tag{32.3}$$

Assume that after performing v_i units, the process is interrupted for a period of length d_i, during which, if there had been no interruption, an additional r_i repetitions would have been produced. In any cycle i, $r_i \leq R_i$ when $d_i \leq D$. Then from Equation 32.2 the time required to perform repetition number $v_i + r_i$ by the end of the break period on the forget curve is:

$$\hat{T}_{v_i + r_i} = \hat{T}_{1i} (v_i + r_i)^f \tag{32.4}$$

Now, denote u_{i+1} as the number of equivalent repetitions remembered at the beginning of cycle $i+1$ after interruptions in i previous cycles. From Equations 32.4 and 32.1, u_{i+1} can be expressed as:

$$u_{i+1} = \left(\frac{\hat{T}_{1i}}{T_1} (v_i + r_i)^f \right)^{-1/b} \tag{32.5}$$

and $\tilde{T}_{1,i+1}$, the time required to perform the first repetition in cycle $i+1$, as:

$$\hat{T}_{v_i+R_i} = \hat{T}_{1i} (v_i + R_i)^f \tag{32.6}$$

32.3.3 Learn–Forget Curve Model (LFCM)

The LFCM (Jaber and Bonney, 1996) allows both the intercept and the forgetting slope to vary from cycle to cycle; i.e. $\hat{T}_1 = \hat{T}_{11} \neq \hat{T}_{12} \neq \hat{T}_{13} \neq \ldots \neq \hat{T}_{1i}$ and $f_1 \neq f_2 \neq f_3 \neq \ldots \neq f_i$ for every $i = 1, 2, 3, \ldots$ The forgetting rate is a function of several parameters and variables, and it is represented as:

$$f_i = \frac{b(1-b)\log v_i}{\log\left(1 + D/t(v_i)\right)} \tag{32.7}$$

where $t(v_i)$ is the time required to perform v_1 repetitions, and it is expressed as:

$$t(v_i) = \begin{cases} \dfrac{T_1}{1-b} v_i^{1-b} & \text{, if } v_i \leq n_s \\ \dfrac{T_1}{1-b} n_s^{1-b} + T_s (v_i - n_s), & \text{otherwise} \end{cases} \tag{32.8}$$

For each interruption, the forget curve intercept is determined by:

$$\hat{T}_{1i} = T_1 v_i^{-(b+f_i)} \tag{32.9}$$

The amount of units remembered at the beginning of cycle $i+1$ is given as:

$$u_{i+1} = \left(\frac{\hat{T}_{1i}}{T_1} (v_i + r_i)^{f_i} \right)^{-1/b} \tag{32.10}$$

The value of $\hat{T}_{v_i+R_i}$ converges to T_1 when total forgetting is assumed, and is calculated as:

$$\hat{T}_{v_i+R_i} = \hat{T}_{1i} (v_i + R_i)^{f_i} = T_1 \tag{32.11}$$

32.4 Results

Table 32.1 contains a summary of the simulation results for the LFCM and VRVF models. In interpreting the values in Table 32.1 (as well as Tables 32.2 and 32.3) it is important to understand how the simulation would behave in the absence of forgetting. The initial processing time for jobs is 4 time units. With no forgetting, the processing time would plateau at 1.0 time units per job after processing 370 jobs. Then, over the course of a 3000-job simulation, the average processing time would be 1.04 time units per job. Thus, we consider values between 1.04 and 1.06 to represent trivial forgetting losses and we have suppressed these values with a dash (–). Higher values of average processing time denote greater forgetting. We will first address the research questions based on the results for LFCM and then highlight some differences in results between LFCM and VRVF.

32.4.1 LFCM Results

The following questions provided motivation for our simulation experiment:

(A) Will providing a worker upfront training reduces forgetting?
(B) How does the frequency of worker transfers relate to forgetting?
(C) Do the answers to above questions A and B change as the number of tasks for which a worker is trained increases or for different forgetting rates?

TABLE 32.1 Average Processing Time Results for LFCM and VRVF

Number of Tasks Learned	Forgetting Rate	Worker Transfer Policy	LFCM Flexibility Acquisition Policy FAP-0	LFCM FAP-1	LFCM FAP-2	VRVF Flexibility Acquisition Policy FAP-0	VRVF FAP-1	VRVF FAP-2
2	Low (95%)	10	1.19	–	–	–	–	–
		25	–	–	–	–	–	–
		250	–	–	–	–	–	–
	Med (90%)	10	1.83	–	–	1.07	–	–
		25	1.41	–	–	1.07	–	–
		250	–	–	–	–	–	–
	High (85%)	10	2.13	–	–	1.13	–	–
		25	1.69	–	–	1.14	–	–
		250	1.08	1.07	1.08	1.07	1.07	–
3	Low (95%)	10	1.67	–	–	1.07	–	–
		25	1.28	–	–	1.07	–	–
		250	–	–	–	–	–	–
	Med (90%)	10	2.18	1.13	–	1.61	1.32	1.24
		25	1.73	1.14	1.07	1.39	1.25	1.21
		250	1.09	1.08	1.07	1.08	1.08	1.07
	High (85%)	10	2.42	–	1.23	2.20	2.02	1.86
		25	1.96	–	1.21	1.81	1.69	1.59
		250	1.15	–	1.09	1.12	1.12	1.10

Note: Average processing time calculated over approximately 3000 jobs per department. The theoretical minimum average processing time is 1.04. Values from 1.04 to 1.06 have been suppressed with a dash "–".

TABLE 32.2 Comparison of Average Processing Time Results for LFCM Augmented with a Task similarity Coefficient for Two Tasks

	Task		Task	Task	Task	Task	
	[redacted]	$T(2)$	[redacted]	[redacted]	[redacted]	T(1)	T(2)
Task Attributes	[redacted]	v_2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	w_2	[redacted]	[redacted]	[redacted]	[redacted]	B
	[redacted]	x_2	[redacted]	[redacted]	[redacted]	C	[redacted]
	[redacted]	y_2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	z_2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Similarity coefficient	$s=0.0$		$s=0.2$	$s=0.4$	$s=0.6$	$s=0.8$	

Forgetting Rate	Worker Transfer Policy	Flexibility Acquisition Policy			Flexibility Acquisition Policy			Flexibility Acquisition Policy			Flexibility Acquisition Policy			Flexibility Acquisition Policy		
		FAP–0	FAP-1	FAP-2	FAP-0	FAP-1	FAP-2	FAP-0	FAP-1	FAP-2	FAP-0	FAP-1	FAP-2	FAP-0	FAP-1	FAP-2
Low (95%)	10	1.19	–	–	1.09	–	–	–	–	–	–	–	–	–	–	–
	25	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	250	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Med (90%)	10	1.83	–	–	1.71	–	–	1.59	–	–	1.36	–	–	1.10	–	–
	25	1.41	–	–	1.30	–	–	1.18	–	–	1.07	–	–	–	–	–
	250	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
High (85%)	10	2.13	–	–	2.02	–	–	1.88	–	–	1.68	–	–	1.39	–	–
	25	1.69	–	–	1.59	–	–	1.46	–	–	1.29	–	–	1.07	–	–
	250	1.08	1.07	1.08	1.07	–	–	–	–	–	–	–	–	–	–	–

Average processing time calculated over approximately 3000 jobs per department.
The theoretical minimum average processing time is 1.04. Values from 1.04 to 1.06 have been suppressed with a dash "–".

TABLE 32.3 Comparison of Average Processing Time Results for LFCM Augmented with Task similarity Coefficients for Three Tasks

	Task			Task			Task			Task		
	[redacted]	$T(2)$	$T(3)$	[redacted]	[redacted]	$T(3)$	[redacted]	[redacted]	$T(3)$	T(1)	T(2)	[redacted]
Task Attributes	[redacted]	v_2	v_3	[redacted]	[redacted]	v_3	[redacted]	[redacted]	v_3	[redacted]	[redacted]	[redacted]
	[redacted]	w_2	w_3	[redacted]	[redacted]	B	[redacted]	[redacted]	B	[redacted]	B	[redacted]
	[redacted]	x_2	x_3	[redacted]	[redacted]	C	[redacted]	[redacted]	C	C	[redacted]	C
	[redacted]	y_2	y_3	[redacted]	[redacted]	y_3	[redacted]	[redacted]	D	[redacted]	[redacted]	D
	[redacted]	z_2	z_3	[redacted]	[redacted]	z_3	[redacted]	[redacted]	z_3	[redacted]	[redacted]	z_3
Similarity Coefficient		$s_{(1,2)} = 0.0$			$s_{(1,2)} = 0.2$			$s_{(1,2)} = 0.25$			$s_{(1,2)} = 0.0$	
		$s_{(2,3)} = 0.0$			$s_{(2,3)} = 0.2$			$s_{(2,3)} = 0.25$			$s_{(2,3)} = 0.0$	
		$s_{(1,3)} = 0.0$			$s_{(1,3)} = 0.2$			$s_{(1,3)} = 0.25$			$s_{(1,3)} = 0.0$	
		$s_{(1,2,3)} = 0.0$			$s_{(1,2,3)} = 0.0$			$s_{(1,2,3)} = 0.2$			$s_{(1,2,3)} = 0.8$	

Forgetting Rate	Worker Transfer Policy	Flexibility Acquisition Policy			Flexibility Acquisition Policy			Flexibility Acquisition Policy			Flexibility Acquisition Policy		
		FAP–0	FAP-1	FAP-2	FAP-0	FAP-1	FAP-2	FAP-0	FAP-1	FAP-2	FAP-0	FAP-1	FAP-2
Low (95%)	10	1.67	–	–	1.53	–	–	1.30	–	–	–	–	–
	25	1.28	–	–	1.16	–	–	1.07	–	–	–	–	–
	250	–	–	–	–	–	–	–	–	–	–	–	–
Med (90%)	10	2.18	1.13	–	2.07	–	–	1.88	–	–	1.40	–	–
	25	1.73	1.14	1.07	1.63	–	–	1.45	–	–	1.07	–	–
	250	1.09	1.08	1.07	1.07	1.07	–	–	–	–	–	–	–
High (85%)	10	2.42	1.38	1.23	2.33	1.24	1.12	2.16	–	–	1.72	–	–
	25	1.96	1.34	1.21	1.87	1.23	1.12	1.71	1.07	–	1.30	–	–
	250	1.15	1.12	1.09	1.12	1.10	1.08	1.08	1.07	–	–	–	–

Average processing time calculated over approximately 3000 jobs per department.
The theoretical minimum average processing time is 1.04. Values from 1.04 to 1.06 have been suppressed with a dash "–".

Looking at the LFCM portion of Table 32.1 we observe that increasing worker training from FAP-0 to FAP-1 to FAP-2 reduces forgetting losses (Question A). Given a worker transfer policy of 10, FAP-1 is sufficient to reduce forgetting losses for all forgetting rates when the number of tasks learned is two, and with a low forgetting rate when the number of tasks learned is three (Question C). When three tasks are learned and the forgetting rate is medium or high, there is additional benefit to FAP-2, more extensive upfront training (Question C).

We also see that reducing the frequency of worker transfers (by increasing transfer policy from 10 to 25 to 250) reduces forgetting losses (Question B). Given no upfront training (FAP-0), increasing worker transfer policy from 10 to 25 is sufficient to counter forgetting losses for two tasks learned and a low forgetting rate (Question C). For higher forgetting rates and for three tasks learned, there is additional benefit to increasing worker transfer policy to 250 (Question C).

While none of our research questions related to the interplay between upfront training and transfer policy, this is perhaps the most interesting component of our results. We see that for two tasks learned, a low forgetting rate (95%), a transfer policy of 10, and no initial training (FAP-0), forgetting losses are moderate with an average processing time of 1.19. Either increasing the transfer policy (to 25 or 250) or introducing initial training (FAP-1 or FAP-2) will remove essentially all forgetting losses. It is not necessary to both increase the transfer policy and provide upfront training, as either method alone is effective in countering forgetting losses.

As the forgetting rate increases to medium (90%) or high (85%) and/or the number of tasks learned increases to three, the average processing times for LFCM generally increase. The general pattern still holds that adding upfront training or increasing the worker transfer policy reduces the forgetting losses. In the case with three tasks learned and a high forgetting rate, we note that *both* providing upfront training and increasing transfer policy are required to achieve a maximum reduction in forgetting losses.

32.4.2 Observations Regarding LFCM and VRVF Results

We note that for three tasks learned the FAP-0 column of LFCM has larger average processing times than for the corresponding entries of FAP-0 for VRVF. However, the average processing times for the FAP-1 and FAP-2 policies under LFCM are smaller than the corresponding values for the VRVF model. In essence, this means that LFCM associates a greater reduction in forgetting loss with increases in upfront training than does VRVF. We feel that the results of LFCM are appropriate for workers learning very dissimilar tasks. That is, we would expect significant benefits from providing upfront training where workers are performing a combination of very different tasks. On the other hand, if there is less dissimilarity in the combination of tasks we would expect a relatively small benefit to accrue from upfront training. For this reason, we feel that VRVF more accurately captures forgetting for tasks that are less dissimilar.

This raises the interesting issue of the effect of task similarity on forgetting. In the next section we present a method for incorporating task similarity in LFCM.

32.5 Augmenting LFCM to Incorporate Task Similarity

While the LFCM provided significant insights into training and deployment issues, no consideration was given to the concept of task similarity. Conceptually, a worker who is being trained on two or three similar tasks is likely to experience relatively less forgetting as compared to being trained on very dissimilar tasks. Here we develop an extension to LFCM in which forgetting is a function of the degree of task similarity.

To define a measure of task similarity, we view each task as being comprised of a set of attributes. For example if task #1 has the set of attributes A, B and C we represent this as $T(1)=\{A, B, C\}$. The definition of the task similarity coefficient is necessarily dependent upon the number of tasks learned.

32.5.1 Similarity Coefficient for Two Tasks

We define the task similarity coefficient for two tasks as:

$$s = \frac{2 \cdot N\{T(1) \cap T(2)\}}{N\{T(1)\} + N\{T(2)\}} \tag{32.12}$$

where $N(X)$ is the number of attributes in task X. Consider two tasks $T(1)$ and $T(2)$ such that $T(1)=\{A, B, C\}$ and $T(2)=\{D, E, F, G\}$. $T(1)$ and $T(2)$ have no attribute(s) in common, and their similarity coefficient equals zero. Now consider an additional task $T(3)=\{A, C, H, I, J\}$. $T(1)$ and $T(3)$ have attributes A and C in common and their similarity coefficient is (2*2)/(3+5)=0.5.

To utilize similarity coefficients, where two tasks are involved, we modify LFCM by replacing Equation 32.9 with Equation 32.13 given below:

$$\hat{T}_{1i} = \begin{cases} (1-s)T_1 v_i^{-(b+f)} + sT_1 v_i^{-b}, & \text{if } v_i < n_s \\ (1-s)T_s v_i^{-f} + sT_s \quad , & \text{otherwise} \end{cases} \tag{32.13}$$

32.5.2 Similarity Coefficients for Three Tasks

For three tasks, $T(1)$, $T(2)$ and $T(3)$, there is one three-way similarity coefficient and three two-way coefficients. The three-way similarity coefficient is calculated as follows:

$$s_{(1,2,3)} = \frac{3 \cdot N\{T(1) \cap T(2) \cap T(3)\}}{N\{T(1)\} + N\{T(2)\} + N\{T(3)\}} \tag{32.14}$$

The two-way coefficient for tasks $T(1)$ and $T(2)$ is given by:

$$s_{(1,2)} = \frac{2 \cdot N\{T(1)^* \cap T(2)^*\}}{N\{T(1)^*\} + N\{T(2)^*\}} \tag{32.15}$$

where $T(t)^* = T(i) - \{T(1) \cap T(2) \cap T(3)\}$ for $i=1, 2$. The other two two-way coefficients, $s_{(2,3)}$ and $s_{(1,3)}$, are calculated similarly.

Consider tasks $T(1)=\{A, w_1, C, D, z_1\}$, $T(2)=\{A, B, x_2, D, z_2\}$ and $T(3)=\{v_3, B, C, D, z_3\}$. Then, $T(1) \cap T(2) \cap T(3)=\{D\}$, $N\{T(1) \cap T(2) \cap T(3)\}=1$, and $N\{T(1)\}+N\{T(2)\}+N\{T(3)\}=15$ and thus $s_{(1,2,3)}=(3\cdot 1)/15=0.2$.

To calculate the two-way similarity coefficient we note that:

$$T(1)^* = \mathrm{T}(1) - \{T(1) \cap T(2) \cap T(3)\} = \{A, w_1, C, z_1\}$$

$$T(2)^* = \mathrm{T}(2) - \{T(1) \cap T(2) \cap T(3)\} = \{A, B, x_2, z_2\}$$

$$T(3)^* = \mathrm{T}(3) - \{T(1) \cap T(2) \cap T(3)\} = \{v_3, B, C, z_3\}$$

Then,

$$s_{(1,2)} = \frac{2 \cdot N\{T(1)^* \cap T(2)^*\}}{N\{T(1)^*\} + N\{T(2)^*\}} = \frac{2 \cdot 1}{4+4} = 0.25$$

$$s_{(2,3)} = \frac{2 \cdot N\{T(2)^* \cap T(3)^*\}}{N\{T(2)^*\} + N\{T(3)^*\}} = \frac{2 \cdot 1}{4+4} = 0.25 \text{ and}$$

$$s_{(1,3)} = \frac{2 \cdot N\{T(1)^* \cap T(3)^*\}}{N\{T(1)^*\} + N\{T(3)^*\}} = \frac{2 \cdot 1}{4+4} = 0.25 \cdot$$

Then, to include similarity coefficients for three tasks we modify LFCM by replacing Equation 32.9 with Equation 32.16:

$$\hat{T}_{1i} = \begin{cases} (1 - s_{jk} - s_{jkl}) T_1 v_i^{-(b+f)} + (s_{jk} + s_{jkl}) T_1 v_i^{-b}, & \text{if } v_i < n_s \\ (1 - s_{jk} - s_{jkl}) T_s v_i^{-f} + (s_{jk} + s_{jkl}) T_s & \text{, otherwise} \end{cases} \tag{32.16}$$

32.5.3 Results for LFCM Augmented with Similarity Coefficients

To demonstrate the augmented LFCM, we present average processing time results in Table 32.2 for five pairs of tasks with different degrees of similarity. The top portion of Table 32.2 shows pairs of tasks with their attributes and corresponding similarity coefficients. For the first pair of tasks, the similarity coefficient is zero and equivalent to the original LFCM. As the level of similarity increases we see that average processing times decrease. Because forgetting losses decrease with increasing similarity, there is less motivation to provide upfront training or reduce transfer frequency.

Table 32.3 presents results for the augmented LFCM with three tasks. Once again the top portion shows tasks with associated attributes and similarity coefficients. On the far left side, tasks $T(1)$, $T(2)$ and $T(3)$ have no common attributes. With each step to the right, the triplet, $T(1)$, $T(2)$ and $T(3)$, is modified by increasing the commonality of the attributes. When tasks similarity is low and the forgetting rate is high, losses due to forgetting can be significant (see the leftmost shaded average processing time, 2.42). Extensive training (FAP-2) alone cannot completely counter forgetting losses, and limiting worker transfers becomes an important consideration. However, as the degree of similarity increases (see the shaded numbers in the two rightmost panels—2.16 and 1.72), even a moderate level of training (FAP-1) can help counter forgetting losses. Furthermore, it is possible to use greater worker flexibility in these shops. Once again, with increasing similarity, the need to counter forgetting losses through upfront training or transfer policy is reduced.

32.6 Conclusions

Our study contributes in the following ways in studying worker training and deployment policies in the presence of learning and forgetting effects. First, we provide a summary of field studies on human forgetting to identify characteristics that a forgetting model should embody. Second, we evaluate the extent to which existing learning–forgetting models incorporate these characteristics of forgetting.

In simulating worker learning and forgetting effects using LFCM, we find that both upfront training and transfer policy can be effective tools for countering forgetting losses. Under many circumstances it is not necessary to use both these tools. However, when more than two tasks are involved and the forgetting rate is high, *both* upfront training and transfer policy may be required to counter forgetting losses.

We introduce a modification to LFCM to incorporate the degree of task similarity. We find that with increasing similarity, the importance of upfront training and transfer policy decline. This result has important implications for work environments in which extensive worker flexibility is desirable for performing a set of closely related tasks.

Acknowledgment

Dr. M.Y. Jaber thanks the Natural Sciences and Engineering Council of Canada (NSERC) for the support of his research.

Appendix

VRIF (Elmaghraby, 1990) uses a power forgetting function with a single intercept $\hat{T}_1$, and forgetting exponent *f*. These parameters, $\hat{T}_1$ and *f*, are either given *a priori,* or derived on the basis of the first lot produced. Once these values are determined, they remain constant for all production cycles *i*, i.e. $\hat{T}_{1i} = \hat{T}_1$ and $f_i = f_1 = f$.

To illustrate how the VRIF model captures forgetting, we consider a situation where the interruption in learning occurs after v_i jobs have been processed in cycle *i* (where $i=1, 2, 3$, etc), where v_i is the sum of the number of jobs remembered at beginning of cycle *i*, u_i, and those completed in *i*, n_i, i.e. $v_i = u_i + n_i$. The forgetting function is then defined by equating the time required to produce the job number v_i on the learning curve to its equivalent on the forgetting curve. This is done by equating Equations 32.1 and 32.2 (in this chapter), and solving for $\hat{T}_1$ to get:

$$\hat{T}_1 = \begin{cases} T_1 v_1^{-(b+f)}, & \text{if } v_1 < n_s \\ T_s v_1^{-f}, & \text{otherwise} \end{cases} \tag{A.1}$$

Now, assume that after completing v_i units, the worker is interrupted for a period of length d_i during which an additional r_i jobs would have been produced. Then, from Equation 32.2, the equivalent performance on the forgetting curve is computed as:

$$\hat{T}_{v_i+r_i} = \hat{T}_1 (v_i + r_i)^f \tag{A.2}$$

Denote u_{i+1} as the number of equivalent jobs remembered at the beginning of cycle $i+1$ after interruptions in previous *i* cycles. Substituting Equation A.2 in Equation 32.1, and solving for u_{i+1} gives:

$$u_{i+1} = \left(\frac{\hat{T}_1}{T_1} (v_i + r_i)^f \right)^{-1/b} \tag{A.3}$$

Substituting Equation A.3 in Equation 32.1 gives the time required to perform the first job in cycle $i+1$ as:

$$\tilde{T}_{1,i+1} = T_1 (u_{i+1} + 1)^{-b} \tag{A.4}$$

By assuming constant values of $\hat{T}_1$ and f, VRIF provides illogical results whenever one of the following three situations occur:

(1) If the number of equivalent jobs processed, n_i+u_i, in cycle $i=1$ is such that the intercept of the forgetting curve, $\hat{T}_1$, is less than the standard time, T_s, and $n_1+u_1 \le n_s$, where $u_1=0$. Then $\hat{T}_1 = T_1 n_1^{-(f+b)} \le T_s \Rightarrow T_1 n_1^{-(f+b)} \le T_1 n_s^{-b} \Rightarrow n_1 \le n_s^{b/(f+b)}$, and $\hat{T}_1(n_1+r_1)^f \le T_s$. Then the forgetting curve will revert to T_s corresponding to n_s on repetition axis, such that $n_s > n_i+u_i$. This result is illogical since a worker cannot remember more than n_i+u_i. In other words, a worker cannot remember more than what she/he has learned.

(2) If the number of jobs performed in cycle i, x_i, is such that, $T_s < \hat{T}x_i^f < T_1(n_i+u_i)^{-b}$, then the forgetting curve reverts to $\hat{T}x_i^f$ that has no equal on the learning curve, and thus u_{i+1} cannot be computed. This results in a discontinuity in the learning–forgetting process, which is illogical.

(3) Finally, when the number of jobs x_i is such that, $T_1(n_i+u_i)^{-b} < \hat{T}x_i^f$ and $T_1(n'_i+u_i)^{-b} = \hat{T}x_i^f$ (where $n'_i < n_i$), the worker starts forgetting while learning which is illogical since the VRIF, like all existing learning–forgetting models, advocates that forgetting occurs only when production ceases for a period of time, and that in-cycle forgetting does not occur.

Thus, VRIF provides logical results, if and only if, in any cycle i the condition $u_i+n_i=n_1$ holds.

References

Arzi, Y. and Shtub, A. 1997. Learning and forgetting in mental and mechanical tasks: A comparative study. *IIE Transactions,* 29(9), 759–768.

Ash, R. and Smith-Daniels, D.E. 1999. The effects of learning, forgetting and relearning on decision rule performance in multi-project scheduling. *Decision Sciences,* 30(1), 47–82.

Bailey, C.D. 1989. Forgetting and the learning curve: A laboratory study. *Management Science,* 35(3), 340–352.

Bailey, C.D. and McIntyre, E.V. 1997. The relation between fit and prediction for alternative forms of learning curves and relearning curves. *IIE Transactions,* 29(6), 487–495.

Carlson, J.G. and Rowe, R.G. 1976. How much does forgetting cost? *Industrial Engineering,* 8(9), 40–47.

Dar-El, E.M. Ayas, K., and Gilad, I. 1995a. A dual-phase model for the individual learning process in industrial tasks. *IIE Transactions,* 27(3), 265–271.

Dar-El, E.M. Ayas, K., and Gilad, I. 1995b. Predicting performance times for long cycle time tasks. *IIE Transactions,* 27(3), 272–281.

Elmaghraby, S.E. 1990. Economic manufacturing quantities under conditions of learning and forgetting (EMQ/LaF). *Production Planning and Control,* 1(4), 196–208.

Globerson, S., Levin, N., and Shtub, A. 1989. The impact of breaks on forgetting when performing a repetitive task. *IIE Transactions,* 21(4), 376–381.

Hewitt, D., Sprague, K., Yearout, R., Lisnerski, D., and Sparks, C. 1992. The effects of unequal relearning rates on estimating forgetting parameters associated with Performance curves. *International Journal of Industrial Ergonomics,* 10(3), 217–224.

Jaber, M.Y. and Bonney, M.C. 1996. Production breaks and the learning curve: The forgetting phenomena. *Applied Mathematical Modelling,* 20(2), 162–169.

Jaber, M.Y. and Bonney, M.C. 1997. A comparative study of learning curves with forgetting. *Applied Mathematical Modelling,* 21(8), 523–531.

Kher, H.V. 2000. Examination of flexibility acquisition policies in dual resource constrained job shops with simultaneous worker learning and forgetting effects. *Journal of Operational Research Society,* 51(5), 592–601.

Kher, H.V., Malhotra, M.K., Philipoom, P.R. and Fry, T.D. 1999. Modelling simultaneous worker learning and forgetting in dual resource constrained systems. *European Journal of Operational Research,* 115(1), 158–172.

McCreery, J.K. and Krajewski, L.J. 1999. Improving performance using workforce flexibility in an assembly environment with learning and forgetting effects. *International Journal of Production Research,* 37(9), 2031–2058.

Nembhard, D.A. 2000. The effect of task complexity and experience on learning and forgetting: A field study. *Human Factors,* 42(2), 272–286.

Nembhard, D.A., and Uzumeri, M.V. 2000. Experiential learning and forgetting for manual and cognitive tasks. *International Journal of Industrial Ergonomics,* 25(4), 315–326.

Park, P.S. and Bobrowski, P.M. 1989. Job release and labor flexibility in a dual resource constrained job shop. *Journal of Operations Management*, 8(3), 230–249.

Shtub, A., Levin, N., and Globerson, S. 1993. Learning and forgetting industrial skills: experimental model. *International Journal of Human Factors in Manufacturing*, 3(3), 293–305.

Sparks, C. and Yearout, R. 1990. The impact of visual display units used for highly cognitive tasks on learning curve models. *Computers and Industrial Engineering*, 19(1–4), 351–355.

Wisner, J.D. 1996. A study of US machine shops with just-in-time customers. *International Journal of Operations and Production Management*, 16(7), 62–76.

Wisner, J.D. and Pearson, J.N. 1993. An exploratory study of the effects of operator relearning in a dual resource constrained job shop. *Production and Operations Management Journal*, 2(1), 55–68.

Wisner, J.D. and Siferd S.P. 1995. A survey of US manufacturing practices in make-to-order machine shops. *Production and Inventory Management Journal*, 36(1), 1–7.

Wright, T. 1936. Factors affecting the cost of airplanes. *Journal of Aeronautical Sciences*, 3(4), 122–128.

33

Half-Life Theory of Learning Curves

Adedeji B. Badiru
Air Force Institute of Technology

Anota O. Ijaduola
Monmouth College

33.1 Introduction

The military is very much interested in training troops fast, thoroughly, and effectively. Team training is particularly of importance as a systems approach to enhancing military readiness. Thus, prediction of team performance is of great importance in any military system. In military training systems that are subject to variability and complexity of interfaces, advance prediction of performance is useful for designing training programs for efficient knowledge acquisition and sustainable retention of skills. Organizations invest in people, work process, and technology for the purpose of achieving increased and enhanced production capability. The systems nature of such investment strategy requires that the investment be a carefully planned activity stretching over multiple years. Learning curve analysis is one method through which system enhancement can be achieved in terms of cost, time, and performance vis-à-vis the strategic investment of funds and other assets. The predictive capability of learning curves is helpful in planning for system performance enhancement and resilience.

Formal analysis of learning curves first emerged in the mid-1930s in connection with the analysis of the production of airplanes (Wright 1936). Learning refers to the improved operational efficiency and cost reduction obtained from repetition of a task. This has a huge impact for training purposes and the design of work. Workers learn and improve by repeating operations. Thus, a system's performance and resilience are dependent on the learning characteristics of its components; with workers being a major component of the system. Learning is time-dependent and externally controllable. The antithesis of learning is *forgetting*. Thus, as learning curve leads to increasing performance through cost reduction, forgetting tends to diminish performance. Considering the diminishing impact of forgetting, the half-life measure will be of interest for assessing the resilience and sustainability of a system. Derivation of the half-life equations of learning curves can reveal more about the properties of the various curves that have been reported in the literature. This chapter presents half-life derivations for some of the classical learning curve models available in the literature.

Several research and application studies have confirmed that human performance improves with reinforcement or frequent and consistent repetitions. Badiru (1992, 1994) provides computational survey of learning curves as well as industrial application to productivity and performance analysis. Reductions in operation processing times achieved through learning curves can directly translate to cost savings. The large wealth of literature on learning curves show that they are referred to by several names, including progress function, cost–quantity relationship, cost curve, production acceleration curve, performance curve, efficiency curve, improvement curve, and learning function. In all of these different perspectives, a primary interest is whether or not a level of learning, once achieved, can be sustained. Sustainability of learning curve is influenced by several factors such as natural degradation, forgetting, and reduction due to work interruption. Thus, it is of interest to predict the future state and behavior of learning curves. In systems planning and control, prediction of performance is useful for determining the line of corrective action that should be taken. Learning curves are used extensively in business, science, technology, engineering, and industry to predict performance over time. Thus, there has been a big interest in the behavior of learning curves over the past several decades.

This chapter introduces the concept of half-life of learning curves as a predictive measure of system performance. Half-life is the amount of time it takes for a quantity to diminish to half of its original size through natural processes. The quantity of interest may be cost, time, performance, skill, throughput, or productivity. Duality is of natural interest in many real-world processes. We often speak of "twice as much" and "half as much" as benchmarks for process analysis. In economic and financial principles, the "rule of 72" refers to the length of time required for an investment to double in value. These common "double" or "half" concepts provide the motivation for the proposed half-life analysis.

The usual application of half-life is in natural sciences. For example, in physics, the half-life is a measure of the stability of a radioactive substance. In practical terms, the half-life attribute of a substance is the time it takes for one-half of the atoms in an initial magnitude to disintegrate. The longer the half-life of a substance, the more stable it is. This provides a good analogy for modeling learning curves with the recognition of increasing performance or decreasing cost with respect to the passage of time. The approach provides another perspective to the large body of literature on learning curves. It can have application not only in the traditional production environment, but also in functions such as system maintenance, safety, security skills, marketing effectiveness, sports skills, cognitive skills, and resilience engineering. The positive impacts of learning curves can be assessed in terms of cost improvement, reduction in production time, or increase in throughput with respect to time. The adverse impacts of forgetting can be assessed in terms of declining performance. We propose the following formal definitions:

For learning curves: Half-life is the production level required to reduce cumulative average cost per unit to half of its original size.

For forgetting curve: Half-life is the amount of time it takes for performance to decline to half its original magnitude.

33.2 Literature on Learning Curves

Although there is extensive collection of classical studies of *improvement* due to learning curves, only very limited attention has been paid to performance *degradation* due to the impact of forgetting. Some of the classical works on process improvement due to learning include Smith (1989), Belkaoui (1976, 1986), Nanda (1979), Pegels (1976), Richardson (1978), Towill (1978), Womer (1979, 1981, 1984), Womer and Gulledge (1983), Camm et al. (1987), Liao (1979), McIntyre (1977), Smunt (1986), Sule (1978), and Yelle (1976, 1979, 1983). It is only in recent years that the recognition of "forgetting" curves began to emerge, as can be seen in more recent literature (Badiru 1995; Jaber and Sikstrom 2004; Jaber et al. 2003; Jaber and Bonney 2003; Jaber and Bonney 2007; Jaber and Guiffrida 2008). The new and emerging research on the forgetting components of learning curves provides the motivation for studying half-life properties of learning curves. Performance decay can occur due to several factors, including lack of training, reduced retention of skills, lapsed in performance, extended breaks in practice, and natural forgetting.

33.3 Analysis of System Performance and Resilience

Resilience engineering is an emerging area of systems analysis that relates to the collection of activities designed to create the ability of a system (or organization) to continue to operate under extremely adverse conditions such as a "shock" or attack. Thus, a system's resilience is indicative of the system's level of performance under shock. If the learning characteristic of the system is stable and retainable, then the system is said to be very resilient. The ability to predict systems performance using learning curve analysis provides an additional avenue to develop corrective strategies for managing a system. For example, supposing we are interested in how fast an IT system responds to service requests from clients. We can model the system's performance in terms of its response time with respect to the passage of time. In this case, it is reasonable to expect the system to improve over time because of the positive impact of learning curve of the IT workers. Figure 33.1 shows a graphical representation of the response time as a function of time. The response time decreases as time progresses, thus indicating increasing performance. The shorter the response time, the more resilient we can expect the system to be in the case of a shock to the system. Typical learning curves measure cost or time reduction. But the reduction can be translated to and represented as performance improvement. Consequently, computing the half-life of the system can be used to measure how long it will take the system's response time to reduce to half of its starting value. Figure 33.2 shows a generic profile for the case where the performance metric (e.g. number of requests completed per unit time) increases with respect to the passage of time.

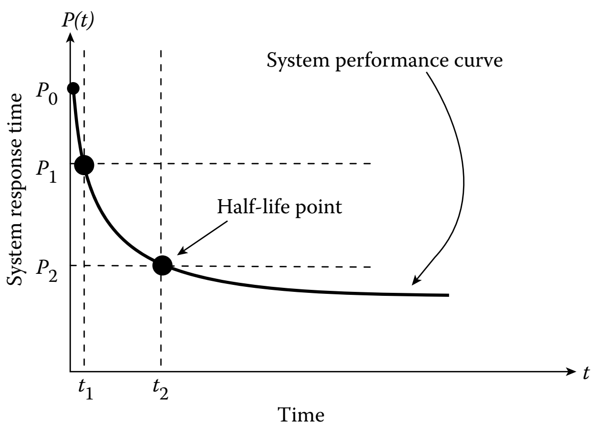


FIGURE 33.1 Representation of system response time with respect to passage of time.

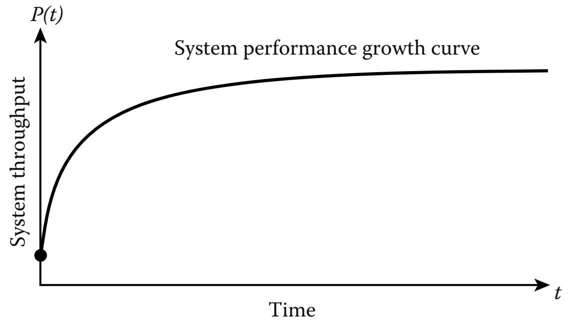


FIGURE 33.2 System performance growth curve.

33.4 Half-life Property of Natural Substances

The half-life concept of learning curve measures the amount of time it takes for performance to degrade by half. Degradation of performance occurs both through natural and imposed processes. The idea of using the half-life approach comes from physics, where half-life is a measure of the stability of a radioactive substance. The longer the half-life of a substance, the more stable it is. By analogy, the longer the half-life of a learning curve model, the more sustainable the fidelity of the learning curve effect. If learning is not very sustainable, then the system will be more vulnerable to the impact of learning curve decline brought on by such random events as system interruptions. To appreciate the impact of half-life computations, consider an engineering reactor that converts the relatively stable uranium 238 into the isotope plutonium 239. After 15 years, it is determined that 0.043% of the initial amount A_0 of the plutonium has disintegrated. We are interested in determining the half-life of the isotope. From physics, the initial value problem is stated as:

$$\frac{dA}{dt} = kA$$

with $A(0) = A_0$. This has a general solution of the form:

$$A(t) = A_0 e^{kt}$$

If 0.043% of the atoms in A_0 has disintegrated, then 99.957% of the substance remains. To find k, we will solve:

$$\alpha A_0 = A_0 e^{15k}$$

where α = remaining fraction of the substance. With $\alpha = 0.99957$we obtain $k = -0.00002867$. Thus, for any time t, the amount of the plutonium isotope remaining is represented as:

$$A(t) = A_0 e^{-0.00002867t}$$

This has a general decay profile similar to the plot of *P(t)* in Figure 33.1. Now, we can compute the half-life as corresponding value at time t for which $A(t) = A_0/2$. That is:

$$\frac{A_0}{2} = A_0 e^{-0.00002867t}$$

which yields *t* (half-life) value of 24,180 years. With this general knowledge of the half-life, several computational analyses can be done to predict the behavior and magnitude of the substance over time. The following examples further illustrate the utility of half-life computations. Let us consider a radioactive nuclide which has a half-life of 30 years. Suppose we are interested in computing the fraction of an initially pure sample of this nuclide that will remain un-decayed at the end of a time period, say 90 years. From the equation of half-life, we can solve for *k*:

$$\frac{A_0}{2} = A_0\, e^{-kt_{\text{half-life}}}$$

$$k = \frac{\ln 2}{t_{\text{half-life}}}$$

Which gives $k=0.0231049$. Now, we can use this value of *k* to obtain the fraction we are interested in computing. That is,

$$\frac{A_0}{A} = e^{-(0.0231049)(90)} = 0.125$$

As another example, let us consider a radioactive isotope with a half-life of 140 days. We can compute the number of days it would take for the sample to decay to one-seventh of its initial magnitude. Thus, we:

$$\frac{A_0}{2} = A_0\, e^{-kt_{\text{half-life}}}$$

$$k = \frac{\ln 2}{t_{\text{half-life}}}$$

Which yields $k=0.004951$. Now, using the value of *k* obtained above, we need to find the time for

$$A = 1/7A_0.$$

That is:

$$\frac{1}{7}A_0 = A_0\, e^{-kt}$$

$$t = \frac{\ln 7}{k} = 393\,\text{days}$$

For learning curves, analogous computations can be used to predict future system performance level and conduct diagnostic assessment of previous performance levels given a present observed performance level. Since there are many alternate models of learning curves, each one can be analyzed to determine its half-life. Thus, a comparative analysis of the different models can be conducted. This general mathematical approach can become the de-facto approach for computational testing of learning curve models.

33.5 Half-life Application to Learning Curves

Learning curves present the relationship between cost (or time) and level of activity on the basis of the effect of learning. An early study by Wright (1936) disclosed the "80% learning" effect, which indicates that a given operation is subject to a 20% productivity improvement each time the activity level or production volume *doubles*. The proposed half-life approach is the antithesis of the double-level milestone. Learning curve can serve as a predictive tool for obtaining time estimates for tasks that are repeated within a project life cycle. A new learning curve does not necessarily commence each time a new operation is started, since workers can sometimes transfer previous skills to new operations. The point at which the learning curve begins to flatten depends on the degree of similarity of the new operation to previously performed operations. Typical learning rates that have been encountered in practice range from 70 to 95%. Several alternate models of learning curves have been presented in the literature. Some of the classical models are:

- Log-linear model
- S-curve model
- Stanford-B model
- DeJong's learning formula
- Levy's adaptation function
- Glover's learning formula
- Pegels' exponential function
- Knecht's upturn model
- Yelle's product model

The basic log-linear model is the most popular learning curve model. It expresses a dependent variable (e.g. production cost) in terms of some independent variable (e.g. cumulative production). The model states that the improvement in productivity is constant (i.e. it has a constant slope) as output increases. That is:

$$C(x) = C_1 x^{-b}$$

or

$$\log C(x) = -b(\log \mathrm{x}) + \log C_1$$

where $C(x)$ = cumulative average cost of producing x units; C_1 = cost of the first unit; x = cumulative production unit; b = learning curve exponent.

Notice that the expression for $C(x)$ is practical only for $x > 0$. This makes sense because learning effect cannot realistically kick in until at least one unit ($x \geq 1$) has been produced. For the standard log-linear model, the expression for the learning rate, p is derived by considering two production levels where one level is double the other. For example, given the two levels x_1 and x_2 (where $x_2 = 2x_1$), we have the following expressions:

$$C(x_1) = C_1(x_1)^{-b}$$

$$C(x_2) = C_1(2x_1)^{-b}$$

The percent productivity gain, p, is then computed as:

$$p = \frac{C(x_2)}{C(x_1)} = \frac{C_1(2x_1)^{-b}}{C_1(x_1)^{-b}} = 2^{-b}$$

The performance curve, $P(x)$, shown earlier in Figure 33.1 can now be defined as the reciprocal of the average cost curve, $C(x)$. Thus, we have

$$P(x) = \frac{1}{C(x)},$$

which will have an increasing profile compared to the asymptotically declining cost curve. In terms of practical application, learning to drive is one example where maximum performance can be achieved in relatively short time compared to the half-life of performance. That is, learning is steep, but the performance curve is relatively flat after steady state is achieved. The application of half-life analysis to learning curves can help address questions such as the ones below:

- How fast and how far can system performance be improved?
- What are the limitations to system performance improvement?
- How resilient is a system to shocks and interruptions to its operation?
- Are the performance goals that are set for the system achievable?

33.6 Derivation of Half-life of the Log-linear Model

Figure 33.3 shows a pictorial representation of the basic log-linear model, with the half-life point indicated as $x_{1/2}$. The half-life of the log-linear model is computed as follows: Let: C_0=initial performance level and $C_{1/2}$=performance level at half-life

$$C_0 = C_1 x_0^{-b} \text{ and } C_{1/2} = C_1 x_{1/2}^{-b}$$

But $C_{1/2} = \frac{1}{2} C_0$

Therefore, $C_1 x_{1/2}^{-b} = \frac{1}{2} C_1 x_0^{-b}$, which leads to $x_{1/2}^{-b} = \frac{1}{2} x_0^{-b}$,

Which, by taking the $(-1/b)^{\text{th}}$ exponent of both sides, simplifies to yield the following expression as the general expression for the standard log-linear learning curve model,

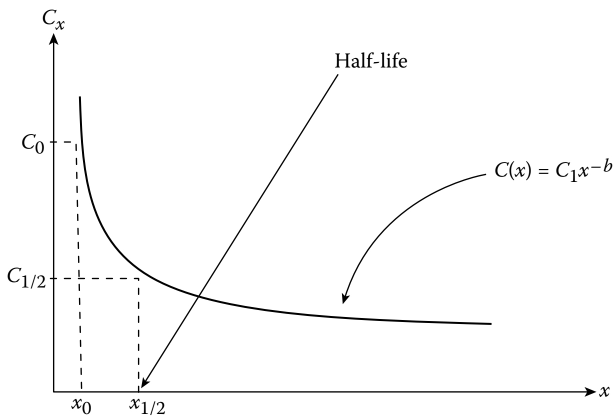


FIGURE 33.3 General profile of the basic learning curve model.

$$x_{1/2} = \left(\frac{1}{2}\right)^{-\frac{1}{b}} x_0, \quad x_0 \geq 1$$

where $x_{1/2}$ is the half-life and x_0 is the initial point of operation. We refer to $x_{1/2}$ (Figure 33.3) as the *first-order half-life*.

The *second-order half-life* is computed as the time corresponding to half of the preceding half. That is:

$$C_1 x_{1/2(2)}^{-b} = \frac{1}{4} C_1 x_0^{-b},$$

which simplifies to yield:

$$x_{1/2(2)} = \left(\frac{1}{2}\right)^{-\frac{2}{b}} x_0,$$

Similarly, the *third-order half-life* is derived to obtain:

$$x_{1/2(3)} = \left(\frac{1}{2}\right)^{-\frac{3}{b}} x_0,$$

In general, the *kth-order half-life* for the log-linear model is represented as:

$$x_{1/2(k)} = \left(\frac{1}{2}\right)^{-\frac{k}{b}} x_0,$$

The characteristics of half-life computations are illustrated in Figures 33.4 and 33.5.

33.7 Computational Examples

Figures 33.2 and 33.3 show examples of log-linear learning curve profiles with $b=0.75$ and $b=0.3032$, respectively. The graphical profiles reveal the characteristics of learning, which can dictate the half-life behavior of the overall learning process. Knowing the point where the half-life of each curve occurs can be very useful in assessing learning retention for the purpose of designing training programs or designing work.

For Figure 33.4 ($C(x)=250x^{-0.75}$), the first-order half-life is computed as:

$$x_{1/2} = \left(\frac{1}{2}\right)^{-\frac{1}{0.75}} x_0, \quad x_0 \geq 1$$

If the above expression is evaluated for $x_0=2$, the first-order half-life yields $x_{1/2}=5.0397$; which indicates a fast drop in the value of *C(x)*. Table 33.1 shows values of $C(x)$ as a function of the starting point, x_0. The specific case of $x_0=2$ is highlighted in Table 33.1. It shows $C(2)=148.6509$ corresponding to a half-life of 5.0397. Note that $C(5.0397)=74.7674$, which is about half of 148.6509. The arrows in the table show how

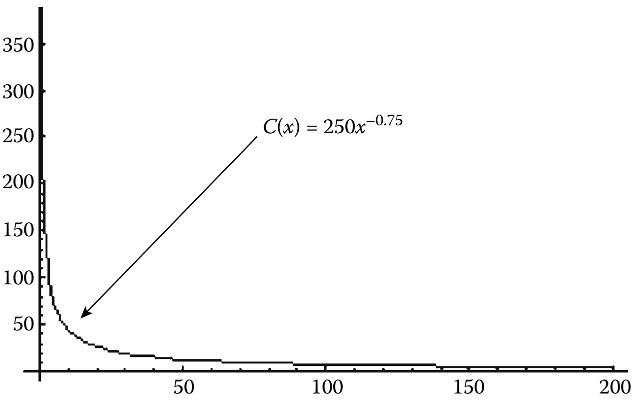


FIGURE 33.4 Learning curve with $b=-0.75$.

TABLE 33.1 Numeric Calculation of Half-lives for $C(x) = 250x^{-0.75}$

x_0	$C(x_0)$	$x_{1/2}$
1	250	2.519842
2	148.6508894	**5.039684**
3	109.6728344	7.559526
4	88.38834765	10.07937
5	**74.76743906**	12.59921
10	44.45698525	25.19842
15	32.79982785	37.79763
20	26.43428159	50.39684
25	22.36067977	62.99605
30	19.50289433	75.59526
35	17.37356628	88.19447
40	15.71791787	100.7937
45	14.38900036	113.3929
50	13.29573974	125.9921
55	12.37849916	138.5913
60	11.59649035	151.1905
65	10.92081352	163.7897
70	10.33038432	176.3889
100	7.90569415	251.9842
120	6.895314416	302.3811
150	5.832725853	377.9763

the various values are linked. The conclusion from this analysis is that if we are operating at the point $x=2$, we can expect this particular curve to reach its half-life decline point at $x=5$.

For Figure 33.5 ($C(x)=240.03x^{-0.3032}$), the first-order half-life is computed as:

$$x_{1/2} = \left(\frac{1}{2}\right)^{-\frac{1}{0.3032}} x_0, \quad x_0 \geq 1$$

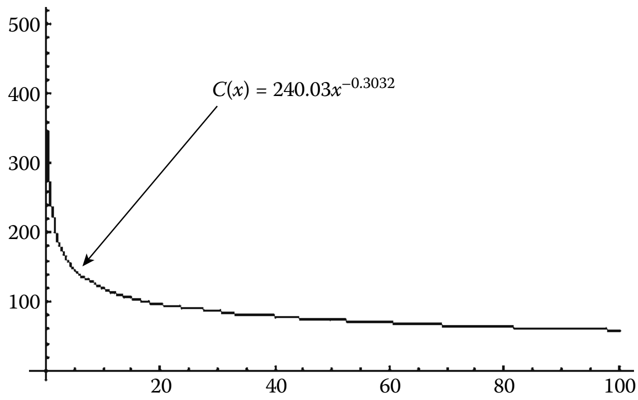


FIGURE 33.5 Learning curve with $b=-0.3032$.

If we evaluate the above function for $x_0=2$; the first-order half-life yields $x_{1/2}=19.6731$. This does not represent as precipitous drop as in Figure 33.4. These numeric examples agree with the projected profiles of the curves in Figures 33.2 and 33.3.

33.8 Cost Expressions for the Log-linear Model

For the log-linear model, using the basic expression for cumulative average cost, the *total cost (TC)* of producing x units is computed as:

$$TC(x) = (x)C_x = (x)C_1x^{-b}$$
$$= C_1x^{(-b+1)}$$

The *unit cost (UC)* of producing the xth unit is given by:

$$UC(x) = C_1x^{(-b+1)} - C_1(x-1)^{(-b+1)}$$
$$= C_1[x^{-b+1} - (x-1)^{-b+1}]$$

The *marginal cost (MC)* of producing the xth unit is given by:

$$MC(x) = \frac{\mathrm{d}[TC_x]}{\mathrm{d}x} = (-b+1)C_1x^{-b}$$

If desired, one can derive half-life expressions for the cost expressions above. For now, we will defer those derivations for interested readers. An important application of learning curve analysis is the calculation of expected production time as illustrated by the following examples. Suppose in a production run of a complex technology component, it was observed that the cumulative hours required to produce 100 units is 100,000 hours with a learning curve effect of 85%. For future project planning purposes, an analyst needs to calculate the number of hours required to produce the 50th unit. Following the standard computations, we have the following: $p=0.85$ and $x=100$ units. Thus, $0.85=2^{-b}$ which yields $b=0.1324$. Consequently, we have $1000=C_1(100)^{-b}$, which yields $C_1=2{,}944.42$ hours. Since b and C_1 are now known, we can compute the cumulative *average* hours required to produce 49 units and 50 units, respectively to obtain $C(49)=1182.09$ hours and $C(50)=1176.50$ hours. Consequently, the *total* hours required to produce the 50th unit is $50[C(50)]-49[C(49)]=902.59$ (approximately 113 work days). If are interested in knowing when these performance metrics would reach half of their original levels in terms of production quantity, we would use half-life calculations.

33.9 Alternate Formulation of the Log-linear Model

An alternate formulation for the log-linear model is called the Unit Cost Model, which is expressed in terms of the specific cost of producing the xth unit, instead of the conventional cumulative average cost expression. The unit cost formula specifies that the individual cost per unit will decrease by a constant percentage as cumulative production doubles. The functional form of the unit cost model is the same as for the average cost model except that the interpretations of the terms are different. It is expressed as:

$$UC(x)=C_1x^{-b}$$

where $UC(x)$ = cost of producing the xth unit; C_1 = cost of the first unit; x = cumulative production count; b = the learning curve exponent as discussed previously.

From the unit cost formula, we can derive expressions for the other cost elements. For the discrete case, the total cost of producing x units is given by:

$$TC(x)=\sum_{q=1}^{x} UC_q = C_1\sum_{q=1}^{x} q^{-b}$$

and the cumulative average cost per unit is given by:

$$C(x)=\frac{C_1}{x}\sum_{q=1}^{x} q^{-b}$$

The marginal cost is found as follows:

$$MC(x)=\frac{\mathrm{d}[TC(x)]}{\mathrm{d}x}=\frac{\mathrm{d}\left[C_1\sum_{i=1}^{x}(i)^{-b}\right]}{\mathrm{d}x}=C_1\frac{\mathrm{d}\left[1+2^{-b}+3^{-b}+\cdots+x^{-b}\right]}{\mathrm{d}x}=C_1bx^{-b-1}$$

For the continuous case, the corresponding cost expressions are:

$$TC(x)=\int_0^x UC(z)\mathrm{d}z=C_1\int_0^x z^{-b}\mathrm{d}z=\frac{C_1x^{(-b+1)}}{-b+1}$$

$$C(x)=\left(\frac{1}{x}\right)\frac{C_1x^{(-b+1)}}{-b+1}$$

$$MC(x)=\frac{\mathrm{d}[TC(x)]}{\mathrm{d}x}=\frac{\mathrm{d}\left[\frac{C_1x^{(-b+1)}}{-b+1}\right]}{\mathrm{d}x}=C_1x^{-b}$$

As in the previous illustrations, the half-life analysis can be applied to the foregoing expressions to determine when each cost element of interest will decrease to half of its starting value. This information can be useful for product pricing purposes, particularly for technology products which are subject to rapid price reductions due to declining product cost. Several models and variations of learning curves have been reported in the literature (see Badiru 1992; Jaber and Guiffrida 2008). Models are developed through one of the following approaches:

1. Conceptual models
2. Theoretical models
3. Observational models
4. Experimental models
5. Empirical models

33.10 Half-life Analysis of Selected Classical Models

33.10.1 The S-Curve Model

The S-curve (Towill and Cherrington 1994) is based on an assumption of a gradual start-up. The function has the shape of the cumulative normal distribution function for the start-up curve and the shape of an operating characteristics function for the learning curve. The gradual start-up is based on the fact that the early stages of production are typically in a transient state with changes in tooling, methods, materials, design, and even changes in the work force. The basic form of the S-curve function is:

$$C(x) = C_1 + M(x+B)^{-b}$$

$$MC(x) = C_1\left[M + (1-M)(x+B)^{-b}\right]$$

where *C(x)*=learning curve expression; *b*=learning curve exponent; *M(x)*=marginal cost expression; C_1=cost of first unit; *M*=incompressibility factor (a constant); *B*=equivalent experience units (a constant).

Assumptions about at least three out of the four parameters (*M*, *B*, C_1, and *b*) are needed to solve for the fourth one. Using the *C(x)* expression and derivation procedure outlined earlier for the log-linear model, the half-life equation for the S-curve learning model is derived to be:

$$x_{1/2} = (1/2)^{-1/b}\left[\frac{M(x_0+B)^{-b} - C_1}{M}\right]^{-1/b} - B$$

where $x_{1/2}$=half-life expression for the S-curve learning model; x_0=initial point of evaluation of performance on the learning curve.

In terms of practical application of the S-curve, consider when a worker begins learning a new task. The individual is slow initially at the tail end of the S-curve. But the rate of learning increases as time goes on, with additional repetitions. This helps the worker to climb the steep-slope segment of the S-curve very rapidly. At the top of the slope, the worker is classified as being proficient with the learned task. From then on, even if the worker puts much effort into improving upon the task, the resultant learning will not be proportional to the effort expended. The top end of the S-curve is often called the slope of *diminishing returns*. At the top of the S-curve, workers succumb to the effects of *forgetting* and other performance impeding factors. As the work environment continues to change, a worker's level of skill and expertise can become obsolete. This is an excellent reason for the application of half-life computations.

33.10.2 The Stanford-B Model

An early form of learning curve is the Stanford-B model, which is represented as:

$$UC(x) = C_1(x+B)^{-b}$$

where $UC(x)$ = direct cost of producing the xth unit; b = learning curve exponent; C_1 = cost of the first unit when $B=0$; B = slope of the asymptote for the curve; B = constant ($1<B<10$). This is equivalent units of previous experience at the start of the process, which represents the number of units produced prior to first unit acceptance. It is noted that when $B=0$, the Stanford-B model reduces to the conventional log-linear model. Figure 33.6 shows the profile of the Stanford-B model with $B=4.2$ and $b=-0.75$. The general expression for the half-life of the Stanford-B model is derived to be:

$$x_{1/2} = (1/2)^{-1/b}(x_0 + B) - B$$

where $x_{1/2}$ = half-life expression for the Stanford-B learning model; x_0 = initial point of evaluation of performance on the learning curve.

33.10.3 Derivation of Half-life for Badiru's Multi-factor Model

Badiru (1994) presents applications of learning and forgetting curves to productivity and performance analysis. One empirical example presented used production data to develop a predictive model of production throughput. Two data replicates are used for each of 10 selected combinations of cost and time values. Observations were recorded for the number of units representing double production levels. The resulting model has the functional form below and the graphical profile shown in Figure 33.7.

$$C(x) = 298.88x_1^{-0.31}x_2^{-0.13}$$

where $C(x)$ = cumulative production volume; x_1 = cumulative units of Factor 1; x_2 = cumulative units of Factor 2; b_1 = first learning curve exponent = −0.31; b_2 = second learning curve exponent = −0.13.

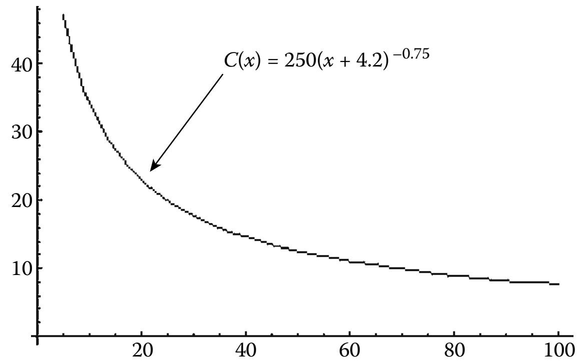


FIGURE 33.6 Stanford-B model with parameters B=4.2 and b=−0.75.

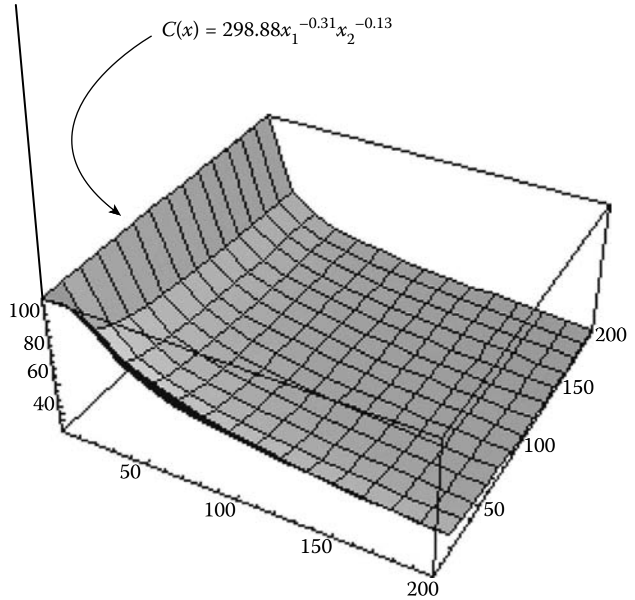


FIGURE 33.7 Bivariate model of learning curve.

A general form of the modeled multifactor learning curve model is:

$$C(x) = C_1 x_1^{-b_1} x_2^{-b_2}$$

and the half-life expression for the multifactor learning curve was derived to be:

$$x_{1(1/2)} = (1/2)^{-1/b_1} \left[\frac{x_{1(0)} x_{2(0)}^{b_2/b_1}}{x_{2(1/2)}^{b_2/b_1}} \right]^{-1/b_1}$$

$$x_{2(1/2)} = (1/2)^{-1/b_2} \left[\frac{x_{2(0)} x_{1(0)}^{b_1/b_2}}{x_{1(1/2)}^{b_2/b_1}} \right]^{-1/b_2}$$

where $x_{i(1/2)}$ = half-life component due to Factor i (i = 1, 2); $x_i(0)$ = initial point of Factor i (i = 1, 2) along the multi-factor learning curve.

Knowledge of the value of one factor is needed to evaluate the other factor. Just as in the case of single-factor models, the half-life analysis of the multi-factor model can be used to predict when the performance metric will reach half of a starting value.

33.10.4 DeJong's Learning Formula

DeJong's learning formula is a power-function which incorporates parameters for the proportion of manual activity in a task. When operations are controlled by manual tasks, the time will be compressible as successive units are completed. If, by contrast, machine cycle times control operations, then the time will be less compressible as the number of units increases. DeJong's formula introduces as incompressible factor, *M*, into the log-linear model to account for the man–machine ratio. The model is expressed as:

$$C(x) = C_1 + Mx^{-b}$$

$$MC(x) = C_1\left[M + (1 - M)x^{-b}\right]$$

where *C(x)* = learning curve expression; *M(x)* = marginal cost expression; *b* = learning curve exponent; *C1* = cost of first unit; *M* = incompressibility factor (a constant).

When *M* = 0, the model reduces to the log-linear model, which implies a completely manual operation. In completely machine-dominated operations, *M* = 1. In that case, the unit cost reduces to a constant equal to C_1, which suggests that no learning-based cost improvement is possible in machine-controlled operations. This represents a condition of high incompressibility. Figure 33.8 shows the profile of DeJong's learning formula for hypothetical parameters of *M* = 0.55 and *b* = –0.75. This profile suggests impracticality at higher values of production. Learning is very steep and average cumulative production cost drops rapidly. The horizontal asymptote for the profile is below the lower bound on the average cost axis, suggesting an infeasible operating region as production volume gets high.

The analysis above agrees with the fact that no significant published data is available on whether or not DeJong's learning formula has been successfully used to account for the degree of automation in any given operation. Using the expression, *MC(x),* the marginal cost half-life of the DeJong's learning model is derived to be:

$$x_{1/2} = (1/2)^{-1/b}\left[\frac{(1 - M)x_0^{-b} - M}{2(1 - M)}\right]^{-1/b}$$

where $x_{1/2}$ = half-life expression for the DeJong's learning curve marginal cost model; x_0 = initial point of evaluation of performance on the marginal cost curve.

If the *C*(*x*) model is used to derive the half-life, then we obtain the following derivation:

$$x_{1/2} = (1/2)^{-1/b}\left[\frac{Mx_0^{-b} - C_1}{M}\right]^{-1/b}$$

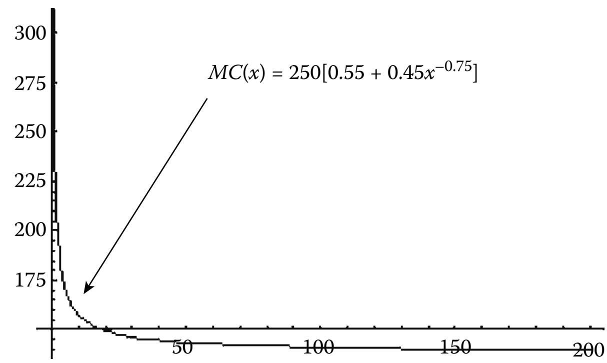


FIGURE 33.8 DeJong's learning formula with *M* = 0.55 and *b* = –0.75.

where $x_{1/2}$ = half-life expression for the DeJong's learning curve model; x_0 = initial point of evaluation of performance on the DeJong's learning curve.

33.10.5 Levy's Adaptation Function

Recognizing that the log-linear model does not account for leveling off of production rate and the factors that may influence learning, Levy (1965) presented the following learning cost function:

$$MC(x) = \left[\frac{1}{\beta} - \left(\frac{1}{\beta} - \frac{x^{-b}}{C_1}\right)k^{-kx}\right]^{-1}$$

where β = production index for the first unit; k = constant used to flatten the learning curve for large values of x.

The flattening constant, k, forces the curve to reach a plateau instead of continuing to decrease or turning in the upward direction. Figure 33.9 shows alternate profiles of Levy's adaptation function

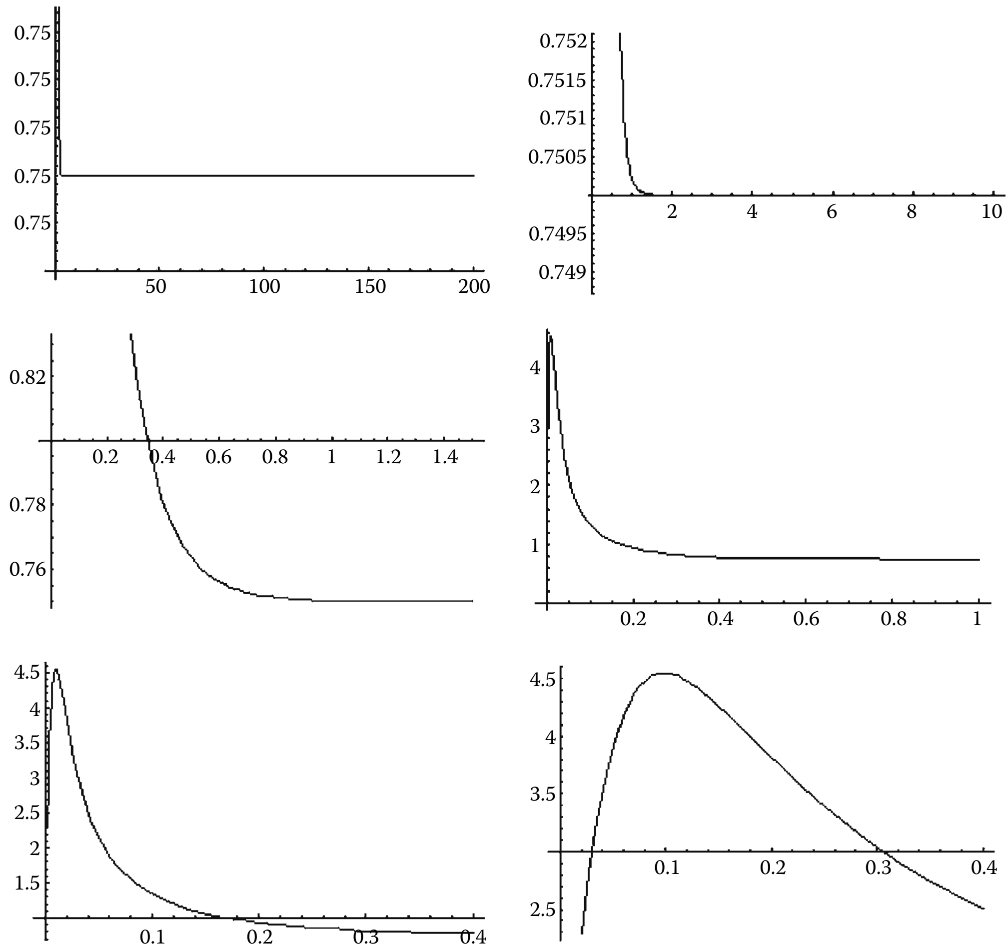


FIGURE 33.9 Profiles of Levy's adaptation over different production ranges.

over different ranges of production for $\beta = 0.75$, $k = 5$, $C_1 = 250$, and $b = 0.75$. The profiles are arranged in increasing order of ranges of operating intervals. The half-life expression for Levy's learning model is a complex nonlinear expression derived as shown below:

$$(1/\beta - x_{1/2}^{-b}/C_1)k^{-kx_{1/2}} = 1/\beta - 2[1/\beta - (1/\beta - x_0^{-b}/C_1)k^{-kx_0}]$$

where $x_{1/2}$ = half-life expression for the Levy's learning curve model; x_0 = initial point of evaluation of performance on the Levy's learning curve.

Knowledge of some of the parameters of the model is needed to solve for the half-life as a closed form expression.

33.10.6 Glover's Learning Model

Glover's learning formula (Glover 1966) is a learning curve model that incorporates a work commencement factor. The model is based on a bottom-up approach which uses individual worker learning results as the basis for plant-wide learning curve standards. The functional form of the model is expressed as:

$$\sum_{i=1}^{n} y_i + a = C_1\left(\sum_{i=1}^{n} x_i\right)^m$$

where y_i = elapsed time or cumulative quantity; x_i = cumulative quantity or elapsed time; a = commencement factor; n = index of the curve (usually $1 + b$); m = model parameter.

This is a complex expression for which half-life expression is not easily computable. We defer the half-life analysis of Levy's learning curve model for further research by interested readers.

33.10.7 Pegel's Exponential Function

Pegels (1976) presented an alternate algebraic function for the learning curve. His model, a form of an exponential function of marginal cost, is represented as:

$$MC(x) = \alpha a^{x-1} + \beta$$

where α, β, and a are parameters based on empirical data analysis. The total cost of producing x units is derived from the marginal cost as follows:

$$TC(x) = \int \left(\alpha a^{x-1} + \beta\right) \mathrm{d}x = \frac{\alpha a^{x-1}}{\ln(a)} + \beta x + c$$

where c is a constant to be derived after the other parameters are found. The constant can be found by letting the marginal cost, total cost, and average cost of the first unit to be all equal. That is, $MC_1 = TC_1 = AC_1$, which yields:

$$c = \alpha - \frac{\alpha}{\ln(a)}$$

The model assumes that the marginal cost of the first unit is known. Thus,

$$MC_1 = \alpha + \beta = y_0$$

Pegel's also presented another mathematical expression for the total labor cost in start-up curves that is expressed as:

$$TC(x) = \frac{a}{1-b} x^{1-b}$$

where x=cumulative number of units produced; a, b=empirically determined parameters.

The expressions for marginal cost, average cost, and unit cost can be derived as shown earlier for other models. Figure 33.10 shows alternate profiles of Pegels' exponential function for α=0.5, β=125, and a=1.2 The functions seem to suggest an unstable process, probably because the hypothetical parameters are incongruent with the empirical range for which the model was developed.

Using the total cost expression, $TC(x)$, we derive the expression for the half-life of Pegel's learning curve model to be as shown below:

$$x_{1/2} = (1/2)^{-1/(1-b)} x_0$$

33.10.8 Knecht's Upturn Model

Knecht (1974) presents a modification to the functional form of the learning curve to analytically express the observed divergence of actual costs from those predicted by learning curve theory when units produced exceed 200. This permits the consideration of nonconstant slopes for the learning curve model. If UC_x is defined as the unit cost of the xth unit, then it approaches 0 asymptotically as x increases. To avoid a zero limit unit cost, the basic functional form is modified. In the continuous case, the formula for cumulative average costs is derived as:

$$C(x) = \int_0^x C_1 z^b \mathrm{d}z = \frac{C_1 x^{b+1}}{(1+b)}$$

This cumulative cost also approaches zero as x goes to infinity. Knecht alters the expression for the cumulative curve to allow for an upturn in the learning curve at large cumulative production levels. He suggested the functional form below:

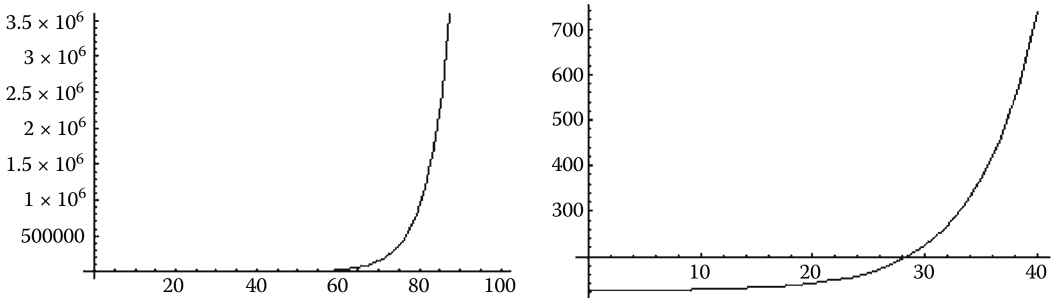


FIGURE 33.10 Alternate forms of Pegels' exponential function for α = 0.5, β = 125 and a = 1.2.

$$C(x) = C_1 x^{-b} e^{cx}$$

where c is a second constant. Differentiating the modified cumulative average cost expression gives the unit cost of the xth unit as shown below. Figure 33.11 shows the cumulative average cost plot of Knecht's function for $C_1 = 250$, $b = 0.25$, and $c = 0.25$.

$$UC(x) = \frac{d}{dx}\left[C_1 x^{-b} e^{cx}\right] = C_1 x^{-b} e^{cx}\left(c + \frac{-b}{x}\right).$$

The half-life expression for Knecht's learning model turns out to be a nonlinear complex function as shown below:

$$x_{1/2} e^{-cx_{1/2}/b} = (1/2)^{-1/b} e^{-cx_0/b} x_0$$

where $x_{1/2}$ = half-life expression for the Knecht's learning curve model; x_0 = initial point of evaluation of performance on the Knecht's learning curve.

Given that x_0 is known, iterative, interpolation, or numerical methods may be needed to solve for the half-life value.

33.10.9 Yelle's Combined Product Learning Curve

Yelle (1976) proposed a learning curve model for products by aggregating and extrapolating the individual learning curve of the operations making up a product on a log-linear plot. The model is expressed as shown below:

$$C(x) = k_1 x_1^{-b_1} + k_2 x_2^{-b_2} + \cdots + k_n x_n^{-b_n}$$

where $C(x)$ = cost of producing the xth unit of the product; n = number of operations making up the product; $k_i x_i^{-bi}$ = learning curve for the ith operation.

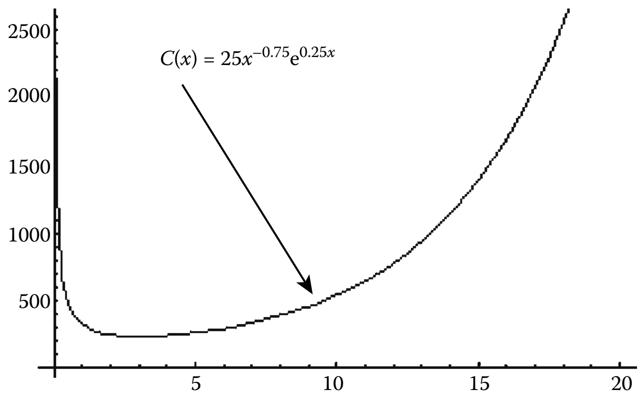


FIGURE 33.11 Knecht's cumulative average cost function for c_1=250, b=0.25, and c=0.25.

The deficiency of Knecht's model above is that a product specific learning curve seems to be a more reasonable model than an integrated product curve. For example, an aggregated learning curve with 96.6% learning rate obtained from individual learning curves with the respective learning rates of 80%, 70%, 85%, 80%, and 85% does not appear to represent reality. If this type of composite improvement is possible, then one can always improve the learning rate for any operation by decomposing it into smaller integrated operations. The additive and multiplicative approaches of reliability functions support the conclusion of impracticality of Knecht's integrated model.

33.11 Contemporary Learning–forgetting Curves

Several factors can influence learning rate in practice. A better understanding of the profiles of learning curves can help in developing forgetting intervention programs and for assessing the sustainability of learning. For example, shifting from learning one operational process to another can influence the half-life profile of the original learning curve. Important questions that half-life analysis can address include the following:

(1) What factors influence learning retention and for how long?
(2) What factors foster forgetting and at what rate?
(3) What joint effects exist to determine the overall learning profile for worker performance and productivity?
(4) What is the profile of and rate of decline of the forgetting curve.

The issues related to the impact of forgetting in performance and productivity analysis are brought to the forefront by Badiru (1994, 1995) and all the references therein. Figure 33.12 shows some of the possible profiles of the forgetting curve. The impact of forgetting can occur continuously over time or discretely over bounded intervals of time. Also, forgetting can occur as random interruptions in the system performance or as scheduled breaks (Anderlohr 1969). The profile of the forgetting curve and its mode of occurrence can influence the half-life measure. This is further evidence that the computation of half-life can help distinguish between learning curves, particularly if a forgetting component is involved.

Recent literature has further highlighted the need to account for the impact of forgetting. Because of the recognition of the diminishing impacts of forgetting curves, these curves are very amenable to the application of the half-life concept. Jaber and Sikstrom (2004) present computational comparisons of three learning and forgetting curves based on previous models available in the literature:

(1) Learn–forget Curve Model (LFCM) provided by Jaber and Bonney (1996).
(2) ReCency Model (RC) provided by Nembhard and Uzumeri (2000).
(3) Power Integration Diffusion (PID) provided by Sikstrom and Jaber (2002).

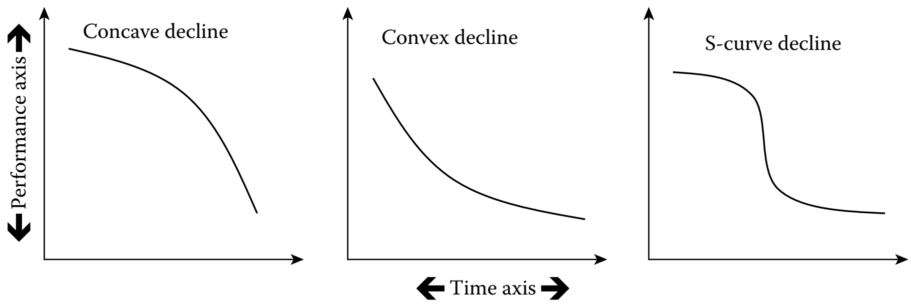


FIGURE 33.12 Alternate profiles declining impact of forgetting.

All three models assume that learning conforms to the original log-linear model presented by Wright (1936) and denoted here as Wright's Learning Curve (WLC):

$$T(x) = T_1 x^{-b},$$

where $T(x)$ is the time to produce the xth unit, T_1 is the time to produce the first unit, x is the cumulative production unit, and b is the learning curve constant $(0 < b < 1)$.

33.11.1 Jaber–Bonney Learn–forget Model

Jaber and Bonney (1996) present the LFCM which suggests that the forgetting curve exponent could be computed as

$$f_i = \frac{b(1-b)\log(u_i + n_i)}{\log(1 + D/t(u_i + n_i))},$$

where $0 \le f_i \le n_i$ is the number of units produced in cycle i up to the point of interruption, D is the break time for which total forgetting occurs, and u_i is the number of units producible due to retained learning at beginning of cycle i from producing x_{i-1} in previous $i - 1$ cycles. Note that in i production cycles, there are $i - 1$ production breaks, where $x_{i-1} = \sum_{j=1}^{i-1} n_j$ and $0 < u_i < x_{i-1}$. That is, if the learning process is interrupted at time of length D, then performance reverts to a threshold value, usually equivalent to T_1. Denote $t(u_i + n_i)$ as the time to produce $u_i + n_i$ units (equivalent units of cumulative production accumulated by the end of cycle i), and b is the learning curve constant. Then, $t(u_i + n_i)$ is computed as presented by Jaber and Sikstrom (2004):

$$t(u_i + n_i) = \sum_{x=1}^{n_i} T_1(u_i + x)^{-b} \cong \int_0^{u_i+n_i} T_1 x^{-b} \mathrm{d}x = \frac{T_1}{1-b}(u_i + n_i)^{1-b}$$

The above function is plotted in Figure 33.13 for $t(u_i + n_i)$ for $T_1 = 25$ and $b = 0.65$.

The number of units produced at the beginning of cycle $i + 1$ is given from Jaber and Bonney (1996) as:

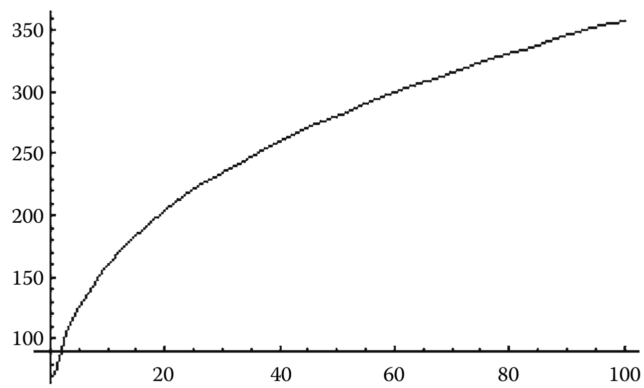


FIGURE 33.13 Plot of Jaber–Sikstrom learn–forget model.

$$u_{i+1} = (u_i + n_i)^{(1+f_i/b)y_i^{-f_i/b}}$$

where $u_1=0$, and y_i is the number of units that would have been accumulated, if production was not ceased for d_i units of time, y_i is computed as:

$$y_i = \left\{ \frac{1-b}{T_1}[t(u_i + n_i) + d_i] \right\}^{1/(1-b)}$$

When total forgetting occurs we have $u_{i+1}=0$. However, $u_{i+1} \rightarrow 0$ as $y_i \rightarrow +\infty$; or alternatively, as $d_i \rightarrow +\infty$, where all the other parameters are of nonzero positive values. Thus, we deduce that total forgetting occurs only when d_i holds a very large value. This does not necessarily contradict the assumption of finite value of D to which total forgetting occurs. Doing so, $u_i+1<1$ when $d_i=D$, and it flattens out at zero for increasing values of $d_i>D$. Anderlohr (1969), McKenna and Glendon (1985), and Globerson et al. (1998) reported findings of the impact of production breaks through empirical studies. As reported by Jaber and Sikstrom (2004), the intercept of the forgetting curve could be determined as:

$$\hat{T}_{1i} = T_1(u_i + n_i)^{-(b+f_i)}$$

The time to produce the first unit in cycle i could then be predicted as

$$\tilde{T}_{1i}^{\text{LFCM}} = T_1(u_i + n_i)^{-b}.$$

33.11.2 Nembhard–Uzumeri Recency (RC) Model

The RC model presented by Nembhard and Uzumeri (2000) has the capability of capturing multiple breaks. Nembhard and Uzumeri (2000) modified the three hyperbolic learning functions of Mazur and Hastie (1978) by introducing a measure termed "recency" of experiential learning, R. For each unit of cumulative production x, Nembhard and Uzumeri (2000) determined the corresponding recency measure, R_x by computing the ratio of the average elapsed time to the elapsed time of the most recent unit produced. Nembhard and Osothsilp (2001) suggested that R_x could be computed as

$$R_x = 2\frac{\sum_{i=1}^{x}(t_i - t_0)}{x(t_x - t_0)}$$

where x is the accumulated number of produced units, t_x is the time when units x is produced, t_0 is the time when the first unit is produced, t_i is the time when unit i is produced, and $R_x \in (1,2)$. The performance of the first unit after a break could be computed as:

$$\tilde{T}_{1i}^{RC} = T_1\left(xR_x^{\alpha}\right)^{-b},$$

where α is a fitted parameter that represents the degree to which the individual forgets the task.

33.11.3 Sikstrom–Jaber Power Integration Diffusion (PID) Model

The PID model presented by Sikstrom and Jaber (2002) advocates that each time a task is performed a memory trace is formed. The strength of this trace decays as a power-function over time. For identical repetitions of a task, an aggregated memory trace could be found by integrating the strength of the memory trace over the time interval of the repeated task. The integral of the power-function memory trace is a power-function. Therefore, the memory strength of an uninterrupted set of repetitions can be described as the difference between a power-function of the retention interval at the start of the repetitions and a power-function of the retention interval at the end of repetitions. The time it takes to perform a task is determined by "a diffusion" process where the strength of the memory constitutes the signal. To simplify the calculation, the noise in the diffusion process is disregarded and the time to perform a task is the inverse of the aggregated memory strength plus a constant reflecting the start time of the diffusion process. The strength of the memory trace follows a power-function of the retention interval since training is given. That is, the strength of a memory trace (at which t time units have passed between learning and forgetting) encoded during a short time interval (dt) is:

$$S'(t) = S_0 t^{-a} \mathrm{d}t$$

where a is the forgetting parameter, $a \in (0,1)$, S_0 is a scaling parameter > 0 (compare with the parameter in other models that represents the time to produce the first unit). The strength of a memory trace encoded for an extended time period is $S(t_{e,1}, t_{e,2})$, where $t_{e,1}$ time units passed since the start of encoding of unit e and $t_{e,2}$ time units passed since the end of encoding of unit e and $t_{e,1} > t_{e,2}$. This memory strength can be calculated by the integral over the time of encoding.

$$S = (t_{e,1}, t_{e,2}) = \int_{t_{e,1}}^{t_{e,2}} S'(t)\mathrm{d}t = \frac{S_0}{1-a}\left[t_{e,2}^{1-a} - t_{e,1}^{1-a}\right]$$

The profile of the above function is plotted in Figure 33.14 for hypothetical values of $S_0 = 20$ and $a = 0.35$.

The strength of the memory trace following encoding during N time intervals is the sum over these intervals, and it is determined as:

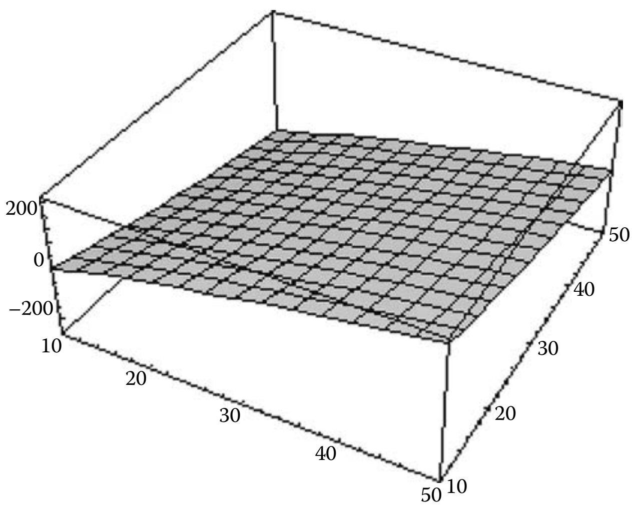


FIGURE 33.14 Plot of Jaber–Sikstrom power integration diffusion model.

$$S=(t_{e,1},t_{e,2})=\frac{S_0}{1-a}\sum_{e=1}^{N}\left[t_{e,2}^{1-a}-t_{e,1}^{1-a}\right]$$

The time to produce a unit is calculated with a diffusion model where the strength of the memory trace is conceived of as a signal. For simplification, the noise in the diffusion process is set to zero. The time to produce a unit is the inverse of the memory strength. The start time of the diffusion process constitutes a constant (t_0) that is added to the total time to produce a unit:

$$\begin{aligned} T(t_r) &= S(t_{e,1},t_{e,2})^{-1}+t_0 \\ &= \frac{1-a}{S_0}\left\{\sum_{e=1}^{N}\left[t_{e,2}^{1-a}-t_{e,1}^{1-a}\right]\right\}^{-1}+t_0 \\ &= S_0'\left\{\sum_{e=1}^{N}\left[t_{e,1}^{a'}-t_{e,1}^{a'}\right]\right\}^{-1}+t_0 \end{aligned}$$

where $S_0'=(1-a)/S_0$ is a rescaling of S_0, and $a'=1-a$, $a'\in(0,1)$ is a rescaling of a. The rescaling of the parameters is introduced for convenience to simplify the final expression. Sikstrom and Jaber (2002) showed that without production breaks, the predictions of PID is a good approximation of WLC model. That is, the predictions are identical given that the accumulated time to produce a unit can be approximated as:

$$t(x)=T_1\sum_{n=1}^{x}n^{-b}\approx T_1\int_0^x n^{-b}=T_1x^{1-b}/(1-b)$$

Thus, in this approximation, Wright's original learning curve model is a special case of PID where

$$\begin{aligned} T_x &= \mathrm{d}t(x)/\mathrm{d}x=\left\{[(1+a')S_0']^{1/(1+a')}\right\}\left(x^{-a'/(1+a')}\right)/(1+a') \\ &= T_1x^{-b} \text{ and } t_0=0 \end{aligned}$$

from which Jaber and Sikstrom (2004) deduce the following relationships, between T_1, a and S_0, and a and b respectively, as:

$$T_1=\frac{[(1+a')S_0']^{1/(1+a')}}{1+a'}, \text{where } x=1$$

and

$$b=\frac{a'}{1+a'}, \text{for every } x>1$$

where $0<b<1/2$ for $0<a'<1$, with $a'=1-a$ and $S_0'=(1-a)/S_0$.

The early studies of learning curves did not address the forgetting function. In this case, the contemporary functions that address the impact of forgetting tend to be more robust and more representative of actual production scenarios. These models can be further enhanced by doing half-life analysis on them.

33.12 Potential Half-life Application to Hyperbolic Decline Curves

Over the years, the decline curve technique has been extensively used by the oil industry to evaluate future oil and gas predictions. These predictions are used as the basis for economic analysis to support development, property sale or purchase, industrial loan provisions, and also to determine if a secondary recovery project should be carried out. It is expected that the profile of hyperbolic decline curve can be adapted for application to learning curve analysis. The graphical solution of the hyperbolic equation is through the use of a log-log paper which sometimes provides a straight line that can be extrapolated for a useful length of time to predict future production levels. This technique, however, sometimes failed to produce the straight line needed for extrapolation for some production scenarios. Furthermore, the graphical method usually involves some manipulation of data, such as shifting, correcting and/or adjusting scales, which eventually introduce bias into the actual data.

In order to avoid the noted graphical problems of hyperbolic decline curves and to accurately predict future performance of a producing well, a nonlinear least-squares technique is often considered. This method does not require any straight line extrapolation for future predictions. The mathematical analysis proceeds as follows: The general hyperbolic decline equation for oil production rate (q) as a function of time (t) can be represented as

$$q(t)=q_0(1+mD_0t)^{-1/m}$$

$$0<m<1$$

where $q(t)$ = oil production at time t; q_0 = initial oil production; D_0 = initial decline; m = decline exponent.

Also, the cumulative oil production at time t, $Q(t)$ can be written as

$$Q(t)=\frac{q_0}{(m-1)D_0}\left[(1+mD_0t)^{\frac{m-1}{m}}-1\right]$$

where $Q(t)$ = cumulative production as of time t.

By combing the above equations and performing some algebraic manipulations, it can be shown that

$$q(t)^{1-m}=q_0{}^{1-m}+(m-1)D_0q_0{}^{-m}Q(t),$$

which shows that the production at time t is a nonlinear function of its cumulative production level. By rewriting the equations in terms of cumulative production, we have

$$Q(t)=\frac{q_0}{(1-m)D_0}+q(t)^{1-m}\frac{q_0^m}{(m-1)D_0}$$

The above function is plotted in Figure 33.15. It can be seen that the model can be investigated both in terms of conventional learning curve techniques, forgetting decline curve, and half-life analysis in a procedure similar to techniques presented earlier in this chapter.

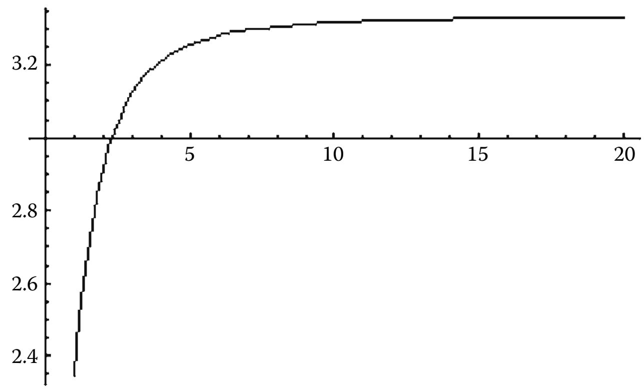


FIGURE 33.15 Plot of hyperbolic decline curve for cumulative production over time.

33.13 Conclusions

Degradation of performance occurs naturally either due to internal processes or externally imposed events, such as extended production breaks. For productivity assessment purposes, it may be of interest to determine the length of time it takes a production metric to decay to half of its original magnitude. For example, for career planning strategy, one may be interested in how long it takes for skills sets to degrade by half in relation to current technological needs of the workplace. The half-life phenomenon may be due to intrinsic factors, such as forgetting, or due to external factors, such as a shift in labor requirements. Half-life analysis can have application in intervention programs designed to achieve reinforcement of learning. It can also have application for assessing the sustainability of skills acquired through training programs. Further research on the theory of half-life of learning curves should be directed to topics such as the following:

- Half-life interpretations
- Learning reinforcement program
- Forgetting intervention and sustainability programs

In addition to the predictive benefits of half-life expressions, they also reveal the ad-hoc nature of some of the classical learning curve models that have been presented in the literature. We recommend that future efforts to develop learning curve models should also attempt to develop the corresponding half-life expressions to provide full operating characteristics of the models. Readers are encouraged to explore half-life analysis of other learning curve models not covered in this chapter.

References

Anderlohr, G. 1969. What production breaks cost. *Industrial Engineering*, 20, 34–36.

Badiru, A.B. 1992. Computational survey of univariate and multivariate learning curve models. *IEEE Transactions on Engineering Management*, 39(2), 176–188.

Badiru, A.B. 1994. Multifactor learning and forgetting models for productivity and performance analysis. *International Journal of Human Factors in Manufacturing*, 4(1), 37–54.

Badiru, A.B. 1995. Multivariate analysis of the effect of learning and forgetting on product quality. *International Journal of Production Research*, 33(3), 777–794.

Belkaoui, A. 1976. Costing through learning. *Cost and Management*, 50(3), 36–40.

Belkaoui, A. 1986. *The Learning Curve*. Quorum Books, Westport, CT.

Camm, J.D., J.R. Evans, and N.K. Womer. 1987. The unit learning curve approximation of total cost. *Computers and Industrial Engineering*, 12(3), 205–213.

Globerson, S., A. Nahumi, and S. Ellis. 1998. Rate of forgetting for motor and cognitive tasks. *International Journal of Cognitive Ergonomics,* 2, 181–191.

Glover, J.H. 1966. Manufacturing progress functions: An alternative model and its comparison with existing functions. *International Journal of Production Research*, 4(4), 279–300.

Jaber, M.Y. and S. Sikstrom. 2004. A numerical comparison of three potential learning and forgetting models. *International Journal of Production Economics*, 92(3), 281–294.

Jaber, M.Y. H.V. Kher, and D. Davis. 2003. Countering forgetting through training and deployment. *International Journal of Production Economics*, 85(1), 33–46.

Jaber, M.Y. and M. Bonney. 1996. Production breaks and the learning curve: The forgetting phenomena. *Applied Mathematical Modelling*, 20, 162–169.

Jaber, M.Y. and M. Bonney. 2003. Lot sizing with learning and forgetting in setups and in product quality. *International Journal of Production Economics*, 83(1), 95–111.

Jaber, M.Y. and M. Bonney. 2007. Economic manufacture quantity (EMQ) model with lot size dependent learning and forgetting rates. *International Journal of Production Economics*, 108(1–2), 359–367.

Jaber, M.Y. and A. L. Guiffrida. 2008. Learning curves for imperfect production processes with reworks and process restoration interruptions. *European Journal of Operational Research*, 189(1), 93–104.

Knecht, G.R. 1974. Costing, technological growth, and generalized learning curves. *Operations Research Quarterly*, 25(3), 487–491.

Levy, F.K. 1965. Adaptation in the production process. *Management Science*, 11(6), B136–B154.

Liao, W.M. 1979. Effects of learning on resource allocation decisions. *Decision Sciences*, 10, 116–125.

Mazur, J.E., and R. Hastie. 1978. Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin*, 85, 1256–1274.

McIntyre, E.V. 1977. Cost-volume-profit analysis adjusted for learning. *Management Science*, 24(2), 149–160.

McKenna, S.P. and A.I. Glendon. 1985. Occupational first aid training: Decay in cardiopulmonary resuscitation (CPR) skills. *Journal of Occupational Psychology*, 58, 109–117.

Nanda, R. 1979. Using learning curves in integration of production resources. Proceedings of 1979 IIE Fall Conference, Chicago, IL, 376–380.

Nembhard, D.A. and N. Osothsilp. 2001. An empirical comparison of forgetting models. *IEEE Transactions on Engineering Management*, 48, 283–291.

Nembhard, D.A. and M.V. Uzumeri. 2000. Experiential learning and forgetting for manual and cognitive tasks. *International Journal of Industrial Ergonomics*, 25, 315–326.

Pegels, C.C. 1976. Start up or learning curves—some new approaches. *Decision Sciences*, 7(4), 705–713.

Richardson, W.J. 1978. Use of learning curves to set goals and monitor progress in cost reduction programs. *Proceedings of 1978 IIE Spring Conference*, St. Louis, MO, 235–239.

Sikstrom, S. and M.Y. Jaber. 2002. The power integration diffusion (PID) model for production breaks. *Journal of Experimental Psychology: Applied*, 8, 118–126.

Smith, J. 1989. *Learning Curve for Cost Control*. Industrial Engineering and Management Press, Norcross, GA.

Smunt, T.L. 1986. A comparison of learning curve analysis and moving average ratio analysis for detailed operational planning. *Decision Sciences*, 17(2), 475–494.

Sule, D.R. 1978. The effect of alternate periods of learning and forgetting on economic manufacturing quantity. *AIIE Transactions*, 10(3), 338–343.

Towill, D.R. and U. Kaloo. 1978. Productivity drift in extended learning curves. *Omega*, 6(4), 295–304.

Towill, D.R. and J.E. Cherrington. 1994. Learning curve models for predicting the performance of advanced manufacturing technology. *International Journal of Advanced Manufacturing Technology*, 9(3), 195–203.

Womer, N.K. 1979. Learning curves, production rate, and program costs. *Management Science***,** 25(4), 312–219.

Womer, N.K. 1981. Some propositions on cost functions. *Southern Economic Journal*, 47, 1111–1119.
Womer, N.K. 1984. Estimating learning curves from aggregate monthly data. *Management Science*, 30(8), 982–992.
Womer, N.K. and T.R. Gulledge, Jr. 1983. A dynamic cost function for an airframe production program. *Engineering Costs and Production Economics*, 7, 213–227.
Wright, T.P. 1936. Factors affecting the cost of airplanes. *Journal of Aeronautical Science*, 3(2), 122–128.
Yelle, L.E. 1976. Estimating learning curves for potential products. *Industrial Marketing Management*, 5(2/3), 147–154.
Yelle, L.E. 1979. The learning curve: Historical review and comprehensive survey. *Decision Sciences*, 10(2), 302–328.
Yelle, L.E. 1983. Adding life cycles to learning curves. *Long Range Planning*, 16(6), 82–87.

34

Readiness for Organizational Change: The Systematic Development of a Scale*

Daniel T. Holt
Air Force Institute of Technology

Achilles A. Armenakis
Auburn University

Hubert S. Feild
Auburn University

Stanley G. Harris
Auburn University

34.1 Introduction

Lewin (1947) argued that during an individual's progression through *change*, the three stages of unfreezing, moving, and refreezing are experienced. Based on this idea, researchers have tried to outline a set of actions that could be taken by change agents to reduce resistance and move organizations and individuals through these stages. For instance, Coch and French (1948) demonstrated the effect various forms of employee participation had on productivity and satisfaction during times of change. They found the greater the extent of participation (i.e. none, participation by representation, and total participation), the more satisfied employees were and the quicker they met new production goals. Others have been spurred by these early efforts to offer further insights into how *resistance* to change could be reduced (cf. Kotter 1995; Kotter and Schlesinger 1979; Lawrence 1954).

Building on this foundation, Armenakis et al. (1993) proposed a model for creating readiness and proposed that readiness was a pre-cursor of resistance and adopti on behaviors. One step in their model was *assessment*. This step is intended to determine just how ready for change employees are before organizational changes are implemented. This assessment enables leaders to identify gaps that may exist between their own expectations about the change initiative and those of other members. If significant gaps are observed and no action taken to close those gaps, resistance would be expected and, therefore, change implementation would be threatened.

* Text of this manuscript previously appeared in Holt, D. T., Armenakis, A. A., Harris, S. G., and Feild, H. S. (2007). Readiness for organizational change: The systematic development of a scale. *Journal of Applied Behavioral Science*, 43(2), 232–255. Published by Sage. Reprinted with permission.

Assessment of readiness can be conducted using both qualitative (e.g. observation and interview techniques) and quantitative (i.e. questionnaire techniques) methods. While qualitative methods provide incredibly rich change-specific information (e.g. Isabella 1990), quantitative methods are an appropriate supplement, offering unique advantages to managers, organizational development consultants, and researchers in certain settings. For instance, a well focused quantitative assessment can be an efficient means to garner change-related information in large global firms because these quantitative instruments can be distributed widely in relatively short periods of time. Furthermore, after a quantitative instrument has been administered, the extent to which the readiness assessment is reliable and valid can be determined. Because of the time and effort that is expended on implementing organizational changes, the reliability and validity of quantitative readiness assessments cannot be overemphasized. Based on this idea, the primary purpose of this manuscript is to propose a quantitative measure of readiness at the individual level that satisfies rigorous psychometric properties (cf. American Psychological Association [APA] 1995), measuring readiness for system-wide changes that impact many facets of organizations.

34.2 Theoretical Framework

Readiness is arguably one of the most important factors involved in employees' initial support for change initiatives (Armenakis et al. 1993, 1999). While the concept of readiness may have been first introduced by Jacobson (1957), the foundation for readiness as a unique construct has been embedded within several theoretical models of the process through which change unfolds. Van de Ven and Poole (1995) synthesized change theories across several disciplines, giving researchers, managers, and organizational development professionals a theoretical means to better understand the phenomenon. Organizational leaders often introduce purposeful, system-wide changes in an effort to realize specified goals (termed teleological change by Van de Ven and Poole). However, as these purposeful changes are introduced, differences and conflicts between the organizational leaders and members may be confronted. For change to occur in the direction that leadership desires, conflicts must be resolved such that organizational members' beliefs and cognitions align with those of the leaders (termed dialectical change by Van de Ven and Poole). In essence, a state of readiness must be created.

Therefore, it is not surprising that the assessment of readiness prior to the introduction of change has been encouraged and several instruments have been developed to fulfill that purpose (Cunningham et al. 2002; Jones et al. 2005; Weeks et al. 2004). These existing instruments appear to measure readiness from one of several perspectives, namely, change process, change content, change context, and individual attributes (Holt et al. 1997). The *change process* refers to the steps followed during implementation. One dimension of change process can be the extent to which employee participation is permitted. A second perspective is the organizational change *content*, which refers to the particular initiative that is being introduced (and its characteristics). Content typically is directed toward administrative, procedural, technological, or structural characteristics of the organization. The third perspective is organizational *context*. Context consists of the conditions and environment within which employees function. For example, a learning organization is one in which employees are likely to embrace continuous change. The fourth and final perspective the *individual attributes* of employees. Because of the differences between individuals, some employees are more inclined to favor organizational changes than others may be.

While it is beyond the scope of this manuscript to offer a comprehensive review of readiness instruments, Holt et al. (2007) reviewed 32 instruments that measure readiness quantitatively. They concluded that there was considerable opportunity for improvement because the available instruments lack evidence of validity and reliability. In sum, two instruments satisfied the standards established by the APA (1995). One, the Lay of the Land Survey (Burke et al. 1996), captured readiness by assessing organizational members' general perceptions of the environment where change was occurring *without* considering a specific initiative. The other, the University of Rhode Island Change Assessment (McConnaughy et al. 1983), did assess readiness for specific initiatives; however, it was designed for changes that were *not organizationally relevant, such as individual efforts to stop smoking or lose weight*. Whereas this instrument has been adapted for use in an organizational setting (Cunningham et al. 2002), it too lacked systematic tests of validity.

Despite the short-comings, Holt et al. (2007) suggest that these instruments have collectively suggested a comprehensive measurement model that is comprised of four factors grounded in the measurement perspectives observed in the existing instruments, namely, the change content, change process, internal context, and individual characteristics. In turn, *readiness for change* was defined as a comprehensive attitude that is influenced simultaneously by the content (i.e., what is being changed), the process (i.e., how the change is being implemented), the context (i.e. circumstances under which the change is occurring), and the individuals (i.e. characteristics of those being asked to change) involved. Furthermore, readiness collectively reflects the extent to which an individual or individuals are cognitively and emotionally inclined to accept, embrace, and adopt a particular plan to purposefully alter the status quo. Table 34.1 depicts the relationship between these four elements and the beliefs among organizational members. Although this model is not explicitly tested in our effort, it does provide a conceptual framework to guide the development of a comprehensive readiness measure, suggesting that a general set of beliefs shape readiness and provide the foundation for resistance or adoptive behaviors.

TABLE 34.1 Results of Factor Analysis

Questionnaire Item (Item numbers based on original questionnaire)	Original Readiness Factor	I	II	III	IV
Factor 1: Appropriateness					
5. I think that the organization will benefit from this change.	OV	0.95	−0.13	−0.02	−.007
4. It doesn't make much sense for us to initiate this change.	D	0.90	−0.07	0.04	−0.03
37. There are legitimate reasons for us to make this change.	D	0.84	−0.11	0.03	−0.19
16. This change will improve our organization's overall efficiency.	OV	0.83	0.00	0.03	0.00
39. There are a number of rational reasons for this change to be made.	D	0.81	0.03	0.01	0.05
1. In the long run, I feel it will be worthwhile for me if the organization adopts this change.	PV	0.80	0.00	−0.29	0.16
7. This change makes my job easier.	PV	0.78	0.12	0.05	−0.10
9. When this change is implemented, I don't believe there is anything for me to gain.	PV	0.77	−0.02	0.05	0.06
32. The time we are spending on this change should be spent on something else.	D	0.71	0.04	0.06	0.09
28. This change matches the priorities of our organization.	OV	0.64	0.17	0.05	0.10
Factor 2: Management support					
42. Our senior leaders have encouraged all of us to embrace this change.	SLS	−.13	0.94	−0.06	0.01
33. Our organization's top decision–makers have put all their support behind this change effort.	SLS	−.07	0.89	−0.01	−0.04
38. Every senior manager has stressed the importance of this change.	SLS	0.05	0.86	−0.10	−0.11
29. This organization's most senior leader is committed to this change.	SLS	0.13	0.68	−0.01	0.06
34. I think we are spending a lot of time on this change when the senior managers don't even want it implemented.	SLS	0.06	0.67	0.02	0.10
6. Management has sent a clear signal this organization is going to change.	SLS	−0.09	0.65	0.19	−0.07
Factor 3: Change efficacy					
20. I do not anticipate any problems adjusting to the work I will have when this change is adopted.	CSE	−0.07	−0.04	0.85	−0.02

(Continued)

TABLE 34.1 Results of Factor Analysis (Continued)

Questionnaire Item (Item numbers based on original questionnaire)	Original Readiness Factor	I	II	III	IV
19. There are some tasks that will be required when we change I don't think I can do well.	CSE	0.11	−0.08	0.78	−0.12
36. When we implement this change, I feel I can handle it with ease.	CSE	−0.05	0.01	0.71	0.06
27. I have the skills that are needed to make this change work.	CSE	−.018	−0.07	0.71	0.27
22. When I set my mind to it, I can learn everything that will be required when this change is adopted.	CSE	0.12	0.14	0.64	−0.07
13. My past experiences make me confident that I will be able to perform successfully after this change is made.	CSE	0.20	0.12	0.51	−0.02
Factor 4: Personally beneficial					
17. I am worried I will lose some of my status in the organization when this change is implemented.	PV	−0.03	−0.08	0.01	0.88
40. This change will disrupt many of the personal relationships I have developed.	PV	−0.05	−0.03	0.02	0.76
15. My future in this job will be limited because of this change.	PV	0.21	0.12	0.06	0.53
Eigenvalues		9.63	2.36	2.07	1.25
Percent of total variance		38.51	9.44	8.26	4.99

Notes: *N*=264. D=discrepancy; OV=organizational valence; PV=personal valence; SLS=senior leadership support; CSE=change self-efficacy.

34.3 Methodological Overview

Based on this, we reasoned that there was an opportunity to build on the insights of the diverse research and develop an *organizationally relevant, change specific instrument*, thereby providing managers, organizational development consultants, and researchers an instrument that might best match their needs. Hinkin (1998) provided a framework to guide the development of a psychometrically sound survey-instrument. The procedure we followed comprises five steps, namely (a) item development, (b) questionnaire administration, (c) item reduction, (d) scale evaluation, and (e) replication with an independent sample. In all, over 900 practicing organizational managers participated in this study. The participants were selected to elicit feedback from a wide range of educational (i.e. participants ranged from high school graduates to those with graduate degrees), functional (e.g. human resource management, engineering, flight operations, and education), and organizational backgrounds (i.e. public- and private-sector). Diversity was emphasized because researchers have suggested that common factors that emerge from heterogeneous samples tend to provide a more general and complete understanding of a phenomenon (e.g. Sutton 1987).

34.3.1 Step 1—Item Development

34.3.1.1 Inductive Development of the Content Domain

Hinkin (1998) suggested that survey items should be developed by specifying the content domain, developing items to assess that domain, and determining the extent to which items measure the specified domain. Because of the many instruments available and the differences observed, available instruments along with 131 books, articles, studies, reports, and papers that addressed the concept of readiness were consulted to inductively refine the content domain of readiness (contact the authors for a list of these documents). To supplement this review, we asked 75 managers from public- as well as private-sector organizations to describe their experiences with recent organizational changes (i.e. critical incidents approach). These descriptions were gathered using two methods. First, a series of semistructured interviews was conducted with senior- to middle-level managers (n = 15). Second, a sample of middle- to lower-level managers (n=60) completed an open-ended questionnaire. Content analysis of the literature, interviews, and open-ended questionnaires yielded 33 themes important to the concept of readiness.

34.3.1.2 Identification of Most Significant Themes

A sample of 291 lower- to mid-level managers reviewed the list of 33 readiness themes that emerged. Participants indicated the extent to which they felt each of these themes impacted an individual's readiness for change, using six-response options (i.e. 1=*extremely negative impact on readiness for change* to 6=*extremely positive impact on readiness for change*). We analyzed the responses and identified the themes that had the highest mean ratings with standard deviations less than or equal to 1.0. We found five themes that satisfied these criteria. The themes were: (a) "confident that you are capable of making the change"; (b) "confident that the change will benefit the employee personally"; (c) "recognition that the organization's leadership supports the change"; (d) "confident that the change will lead to long-term benefits for the organization"; and (e) "recognition of the need for change". We labeled these themes as (a) *self-efficacy*, (b) *personal valence*, (c) *senior leader support*, (d) *organizational valence*, and (e) *discrepancy*. More importantly, these themes aligned with the readiness model presented where content (i.e. organizational valence), process (i.e. management support), context (i.e. discrepancy), and individual attributes (i.e. self-efficacy and personal valence) were represented.

34.3.1.3 Item Development

The five most critical themes, referred to as readiness factors, were formally defined and 59 items that reflected each (11–12 items per factor) were written and evaluated through two formal tests.

Content adequacy test I. The first test was conducted using the procedures detailed by Schriesheim et al. (1993) where participants identified the readiness factors (along with a "none of the above" category) that each of the 59 statements represented. Twenty-six military officers beginning a graduate program in engineering management participated; they were mostly male (n=23) with an average age 26.7 years (SD=3.4).

The data were analyzed using two methods. First, items were retained if 60% of the points assigned were in the appropriate category (cf. Bolino and Turnley 1999). Eighteen of the original 59 items failed to meet this criterion. Three of these 18 items, in fact, did not represent the intended factors. For instance, one item intended to reflect an individual's personal valence ("I am concerned with the risks associated with this change") had only 24.1% of the total points in this category while 27.8% of the points were in the organizational valence category. Perhaps, participants felt the item referred to the organization's risk while others felt the item referred to personal risks.

After the 18 items were eliminated, the data were analyzed using a unique factor analytic approach described by Schrieshiem and his colleagues (1993). Using a conservative standard (cf. Ford et al. 1986), only items that (a) had loadings of at least 0.60 (in absolute terms) on the intended factor and (b) had no other loadings greater than 0.30 (in absolute terms) were considered meaningful. Although all of the items met the loading criterion of 0.60, one item ("This change represents a departure from our organization's values") cross loaded on an unintended factor and was deleted.

Content adequacy test II. Even though the initial test indicated that the majority of the items reflected the intended readiness factors, revisions were made. Specifically, participants had problems distinguishing the items designed to measure discrepancy and organizational valence. Thus, four additional items were written to represent the discrepancy factor rather than modify the factor definitions. These included: (a) "This change is clearly needed"; (b) "There are legitimate reasons for us to make this change"; (c) "I think there are real business needs that make change necessary"; and (d) "No one has explained why this change must be made" (see Table 34.2).

TABLE 34.2 Descriptive Statistics and Correlation Matrix for Sample 1 and Sample 2

	Initial Sample		Correlations														Replication Sample	
Dimension	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	M	SD
Demographic Variables																		
1. Age	47.6	8.7	–	0.12	0.02	−0.03	−0.07	−0.03	−0.06	0.16	–	−0.11	0.02	−0.20*	0.02	0.09	47.01	8.2
2. Gender	0.6	0.5	0.14*	–	0.23*	0.03	−0.13	0.04	−0.12	−0.14	–	−0.09	0.09	−0.13	−0.11	−0.08	0.57	0.5
3. Education	2.6	1.3	0.02	0.25*	–	−0.05	0.15	−0.07	0.05	−0.05	–	0.11	0.12	0.22*	−0.11	0.03	3.62	1.73
4. Organizational Level	2.9	1.6	−0.01	0.03	−0.01	–	−0.19*	0.08	0.05	0.01	–	0.02	0.00	−0.07	−0.04	−0.04	3.00	1.62
Readiness for Change Factors																		
5. Appropriateness	4.5	1.2	−0.08	0.06	0.10	−0.23*	–	0.35*	0.53*	0.49*	–	0.34*	−0.16*	0.24*	0.49*	0.58*	4.0	1.4
6. Management support	5.3	1.1	−0.03	−0.02	−0.04	−0.22*	0.48*	–	0.30*	0.32*	–	0.15*	−0.22*	0.05	0.43*	0.41*	4.9	1.4
7. Change efficacy	5.3	1.0	−0.04	0.08	−0.01	−0.16*	0.47*	0.40*	–	0.53*	–	0.46*	−0.13*	0.27*	0.51*	0.50*	4.8	1.1
8. Personally beneficial	4.9	1.2	−0.06	−0.11	0.03	−0.15*	0.47*	0.42*	0.54*	–	–	0.41*	−0.20*	0.22*	0.46*	0.47*	4.8	1.3
Personality Variables																		
9. Negative affect[a]	1.6	0.5	−0.14*	0.00	−0.06	0.13	−0.07	−0.14*	−0.24*	−0.14*	–	–	–	–	–	–	–	–
10. Locus of control	5.4	0.9	0.00	0.01	0.06	−0.06	0.26*	0.35*	0.46*	0.37*	−0.44*	–	−0.41*	0.46*	0.25*	0.30*	5.2	0.93
11. Rebelliousness	2.9	0.8	−0.19*	0.04	0.00	0.11	−0.19*	−0.36*	−0.29*	−0.37*	0.32*	−0.48*	–	−0.27*	−0.28*	−0.39	2.9	0.83
12. General attitude toward change	4.7	1.3	0.02	−0.13*	0.17*	−0.18*	0.32*	0.22*	0.34*	0.40*	−0.34*	0.49*	−0.44*	–	0.10	0.12	4.3	1.2
Contextual Variables																		
13. Communications climate	4.0	1.2	0.07	−0.07	0.06	−0.12	0.38*	0.48*	0.34*	0.43*	−0.17*	0.49*		−0.44	–	0.58*	4.1	1.15
14. Perceived management ability	4.7	1.3	0.00	−0.09	−0.03	−0.18*	0.55*	0.68*	0.46*	0.50*	−0.19*	0.35*		−0.40	−0.40	–	–	–

Note: Correlation coefficients along the lower diagonal represent those from the first field test ($N=202$–262); those along the upper diagonal are from the replication sample ($N=221$–227).

[a] Negative affect was not measured in the second organization.

*$p<0.05$.

A new sample of 88 judges evaluated this set of 44 items using the approach described by Anderson and Gerbing (1991). The data were analyzed by first computing the proportion of participants that categorized the item as intended. As recommended (Bolino and Turnley 1999; Schriesheim and Hinkin 1990), items with a proportion of agreement greater than 0.70 were retained. Two items designed to tap organizational valence failed to meet this criterion. For instance, one item was correctly categorized by only 68% of the participants ("This change replaces outdated aspects of the organization, while building on the positive attributes of the organization"). Second, *substantive-validity coefficients* (SVC) were computed (Anderson and Gerbing 1991) to statistically assess the extent to which the item assessed another, unintended factor. The SVC of one item that was designed to gauge discrepancy was not statistically significant ($p<0.01$) and eliminated ("I think there are real business needs that make change necessary"). Despite the final deletions, each readiness factor had at least six items remaining and no additional items were written before the questionnaire was administered (Hinkin 1998).

34.3.2 Step 2—Questionnaire Administration

Using the items that were deemed appropriate, a questionnaire was developed that included the readiness items as well as items from known scales. The responses to the readiness items were used in Step 3, Initial item reduction. After Step 3 was complete, the data from the readiness items and known scales were used in Step 4, Scale evaluation (cf. Hinkin 1998). The known scales are described as Step 4 is outlined. All readiness items were phrased in such a way that participants expressed their level of agreement with each item using a seven-point response format ranging from 1 = *strongly disagree* to 7 = *strongly agree*.

34.3.2.1 Organizational Setting

The questionnaire was administered in a government organization that had a $300 million budget and was responsible for developing and fielding information systems for the Department of Defense. The organization's senior leadership with the help of an external consultant had gone through a detailed analysis of key customers' requirements and identified core business functions to support those. Through this analysis, a series of leadership objectives was developed to better align their operations with those requirements and functions. One of these leadership objectives, termed "Organize for success", outlined a new organization structure that clarified lines of authority and eliminated duplicate functions.

34.3.2.2 Sample

Six weeks prior to the implementation of the new structure, 264 employees of this organization (53% response rate) completed the questionnaire. Of these, males represented 59% of the sample, and the age of the average participant was 47.6 years. An array of job titles was represented ranging from illustrator to quality assurance. However, computer analysts and programmers represented the largest portion of the sample. In addition, participants indicated that 2.9 organizational levels, on average, separated their position from the organization's most senior leader.

34.3.3 Step 3—Initial Item Reduction

34.3.3.1 Inter-item Correlations

Hinkin (1998) suggested that scales be refined initially using inter-item correlations and exploratory factor analysis. Items could be eliminated from the initial pool with little or no loss in sampling of the content domain when inter-item correlations between items exceed 0.7, avoiding too much redundancy and artificially inflated estimates of internal consistency (e.g. Boyle 1991). Based on this criterion, four items were deleted.

34.3.3.2 Factor Analysis

Factor analysis was conducted using the methods prescribed by Conway and Huffcutt (2003), Ford et al. (1986), and Hinkin (1998). Thus, the items were factor analyzed using the principle axis method and an orthogonal rotation (cases to item ratio was about 7:1). Six factors initially emerged using the eigenvalue criterion in conjunction with a scree plot; however, 12 items exhibited loadings (factor loading less than 0.4; cross loading greater than 0.35; or factor loading on main factor less than two times that of other factors) that warranted their removal. The remaining items were factor analyzed; factors were extracted, and loadings were evaluated using the same procedure. Four factors emerged accounting for 62.7% of the variance. Table 34.1 shows the results from this analysis along with the original classification of the items (an oblique rotation yielded a similar factor structure, explaining the same amount of variance).

The results for the first factor were more complex than expected. In all, ten items loaded on this factor. Four of the items were intended to measure the extent to which members felt that a change was needed (i.e. discrepancy), representing the participants' perceptions regarding the legitimacy of a change. Three of the items were designed to measure the extent to which members felt the change would be beneficial to the organization (i.e. organizational valence), focusing on the change's benefits, gained efficiency, and goal congruence. The idea that these items tended to cluster onto a single factor was not a complete surprise. The results from the content adequacy tests suggested that participants had problems distinguishing between discrepancy and organizational valence. The results from this factor analysis reinforced this result, indicating that participants in a field setting tend to view these ideas (discrepancy and organizational valence) as a unitary construct. Thus, Factor 1 was labeled *appropriateness*.

Factor 2, termed *management support*, contained six items originally intended to measure this construct. This factor represented the extent to which organizational members felt senior leaders supported the change. Factor 3, termed *change efficacy*, contained six of the original change efficacy items, and reflected the extent to which organizational members felt confident that they would perform well and be successful. The consistency of the items and the magnitude of the loadings provided strong empirical support that these two categories may be influential to an individual's readiness.

The final factor, Factor 4, labeled *personal valence*, included three items originally intended to measure whether the change was perceived to be personally beneficial. Clearly, with only three of the items originally intended to measure this idea loading meaningfully on a distinguishable factor and another item loading on an unintended factor, participants may have a more limited conceptualization of personal valence than we had hypothesized as we developed items. In fact, an examination of the content suggests that this factor reflected concerns about relationships, status, and opportunities while more job related concerns loaded with another factor; a result that was not entirely surprising considering the change under study was a reorganization.

34.3.3.3 Estimates of Internal Consistency

Estimates of internal consistency were computed for each factor. Coefficient alphas were 0.94 for appropriateness, 0.87 for management support, 0.82 for change efficacy, and 0.66 for the personal valence score. While the internal consistency of the personal valence scale did not meet the standard of 0.70 that has been suggested (Nunnally 1978), the standard of 0.70 was relaxed because of the exploratory nature of the scale.

34.3.4 Step 4—Scale Evaluation

Hinkin (1998) recommended that construct and predictive validity of new scales be evaluated beyond the evidence provided through factor analysis. Therefore, known scales designed to measure personality and contextual variables were administered along with the readiness factors so that convergent validity, the extent to which new scales share variance with other known scales, could be explored. Theoretically, variables were included because there appeared to be a theoretical relationship between a particular variable and readiness for change (e.g. Wanberg and Banas 2000). Practically, the variables were selected because reliable and valid measures of these concepts were available. Also, predictive ability of the scales was examined by testing the ability of the readiness factors to (a) distinguish between known groups of participants and nonparticipants; and (b) predict three attitudinal job outcomes commonly studied in change research—job satisfaction, affective commitment, and turnover intentions. Moreover, the ability of the readiness factors to explain incremental variance in predicting these criteria (job satisfaction, affective commitment, and turnover intentions) was tested after controlling for (a) demographic characteristics, (b) personality traits, and (c) culture characteristics.

34.3.4.1 Convergent Validity

Personality factors. Although great care was taken to develop items that would reduce to the intended factors, the inductive scale development procedures we used meant that we were not sure what readiness for change factors would emerge. Not knowing what factors would emerge meant that we could only speculate as to the known scales that should be included to establish some initial level of convergent validity when the readiness for change items were administered in a field setting. However, the recent literature exploring organizational change has suggested a number of personality factors and facets of an organization's culture that could be expected to correlate with readiness for change factors (cf. Wanberg and Banas 2000).

Measures of an individual's locus of control (using the seven-item Internal Mastery Scale developed by Pearlin et al. 1981, $\alpha=0.77$), negative affect (using the 10-item Negative Affect Schedule developed by Watson et al. 1988, $\alpha=0.86$), rebelliousness (using the 11-item scale developed by Hong and Faedda 1996, $\alpha=0.85$), and general attitudes toward change (using the five items developed by Trumbo 1961, $\alpha=0.73$) were included. Locus of control and general attitudes toward change were expected to be positively related to the readiness factors while negative affect and rebelliousness were thought to be negatively related to the readiness factors.

Organizational culture. Consistent with the idea that people's personalities influence readiness, there is a considerable body of literature that suggests that the situation or context also influences readiness. Gopalakrishnan and Damanpour (2000) found certain contextual factors such as size and product scope influenced the speed with which changes were adopted. Others have explored the context by examining the perceptions of the employees who are affected by the changes (Wanberg and Banas 2000). Consistent with this latter notion, we measured the perceptions members had of the organization's communication climate (measured with the four-item scale developed by Miller et al. 1994, $\alpha=0.73$) and perceived ability of management (measured with a six-item scale developed by Mayer and Davis 1999, $\alpha=0.94$). Both of these organizational elements were expected to be positively related to the readiness factors.

Convergent validity results. Means, standard deviations, correlations, and estimates of reliability are provided in Table 34.2. These results indicated that the readiness factors were correlated with each other (mean $r=0.46$, $p<0.05$). In addition, the correlations among the variables gave some evidence of convergent validity. As expected, locus of control and general attitudes toward change were positively related to each of the readiness factors while negative affect and rebelliousness were negatively related to each of the readiness factors. Additionally, perceptions of the communications climate and managements' ability were both positively related to the readiness for change factors.

34.3.4.2 Differences between Known Groups

Participation is generally believed to increase the acceptance of proposed changes. This outcome may occur through a number of mechanisms. First, those who participate in planning and implementing change often have the opportunity to influence the change. Those with this direct influence tend to become affectively committed to the change effort and support the change overtly (Miller and Monge 1985). Second, those who participate often have greater access to change-related information than those who do not. This access to information makes it possible for participants to better understand the justification for change and its ultimate objectives. Therefore, it was reasonable to expect those who actually participated in the development of the new organizational structure to be more ready for change than those who did not participate in the change. That is, participants in the planning of a change initiative should score higher on a valid measure of readiness for change than nonparticipants (cf. Coch and French 1948).

An organization representative identified those who participated in the planning of the change. In all, 50 people were identified. Of these, 43 (86%) participants completed the questionnaire. Participation was a categorical variable coded as a 0 = *nonparticipant* or 1 = *participant*.

Results of known-groups analysis. A one-way MANOVA was conducted to test whether participants reported higher mean readiness than nonparticipants. The results indicated that participation was related to readiness ($F[4,216] = 4.17$, $p < 0.01$). To further explore the differences between participants and nonparticipants on each readiness factor, a series of univariate ANOVAs was conducted. In all cases, participants, as expected, reported higher mean scores on the readiness factors than nonparticipants. Thus, the readiness scales effectively discriminated among groups that were expected to have differing levels of readiness.

34.3.4.3 Predictive and Incremental Validity

Of the 264 employees that completed the first questionnaire, 156 (59%) completed an abbreviated version of the original questionnaire seven months later. This questionnaire included the three-item scale of job satisfaction (Cammann et al. 1983; $\alpha = 0.83$), the six-item scale of affective commitment (Allen and Meyer 1990; $\alpha = 0.86$), and the three-item scale of turnover intentions (Cammann et al. 1983; $\alpha = 0.88$). No significant differences (using *t*-tests) were detected in the mean of organizational level, gender, or age between those who responded to only the Time-1 questionnaire ($n = 108$) and those who responded to both the Time-1 and Time-2 questionnaires ($n = 156$).

Predictive validity results. We hypothesized that the readiness for change factors would be related to job satisfaction, affective commitment, and turnover intentions when these factors were measured well after the change was implemented. Using regression, the readiness for change factors collectively explained 23% ($F = 9.24$, $p < 0.01$), 17% ($F = 8.18$, $p < 0.01$), and 10% ($F = 3.59$, $p < 0.01$) of the variation in the organizational members' job satisfaction, affective commitment, and turnover intentions, respectively. Although each of the readiness for change factors was not related to each of the attitudinal outcomes when all of the readiness factors were included in the regression models (i.e. management support [mean $\beta = 0.11$, $p > 0.05$] and personal valence [mean $\beta = 0.06$, $p > 0.05$] were not related to any of the attitudinal outcome variables), the relationships that did emerge were in the expected directions. That is, appropriateness was positively related to job satisfaction ($\beta = 0.27$, $p < 0.01$) and affective commitment ($\beta = 0.26$, $p < 0.01$), while change efficacy was positively related to job satisfaction ($\beta = 0.25$, $p < 0.01$) and affective commitment ($\beta = 0.21$, $p < 0.05$).

Incremental validity results. In order to test the incremental predictive validity of the readiness factors, hierarchical multiple regression was used. Each of the demographic characteristics, personality variables, and contextual variables were entered first to predict the attitudinal outcomes, then the readiness factors were added to the equations to ascertain any increase in explained variance. This analysis indicated that the addition of the readiness factors in step two increased the explained variance of job satisfaction (when readiness for change factors were entered $\Delta R^2 = 0.08$, $p < 0.05$) and affective commitment (when readiness for change factors were entered $\Delta R^2 = 0.08$, $p < 0.05$). The readiness for change factors did not explain a significant amount of variation in turnover intentions over the demographic characteristics, personality variables, and contextual variables (when readiness for change factors were entered $\Delta R^2 = 0.04$, $p > 0.05$).

34.3.5 Step 5—Replication

While the previous data provided some evidence that the readiness scales were valid and reliable, one study does not establish the validity of a new measure. Therefore, a questionnaire that included the readiness items along with other known scales designed to measure individual characteristics and specific contextual characteristics was administered in another organization.

34.3.5.1 Organizational Setting and Sample

The organization was similar to that in the first field administration in that it was responsible for information technology; however, this second organization was a private-sector firm. Also, the leadership of this organization was implementing a new organizational structure based on the recommendations of an external consultant. While the details of this engagement were not made available to our research team, the change included an additional dimension in that two smaller, geographically separated organizations were being merged into a larger organization. Data were collected from 228 employees (46% response rate). On average, this sample was 47.0 years old.

34.3.5.2 Confirmatory Factor Analysis and Internal Consistency

In this replication study, a confirmatory factor analysis of the readiness items was conducted to further analyze the factor structure and provide additional evidence of the construct validity of the readiness scale. Using a procedure demonstrated by Sturman and Short (2000), the hypothesized four factor model was tested and compared against alternative models that were logical and represented the data with fewer factors. The first alternative was the single factor model, which suggested that the readiness for change items could not be represented by multiple factors or dimensions. Next, a two factor model was tested where the appropriateness, efficacy, and valence items were considered one factor and management support remained as an independent factor. Two, three factor models were tested. In one model, we combined the ten item appropriateness scale with the efficacy scale for one factor and personal valence and management support were considered factors two and three. In the second three factor model, the ten item appropriateness scale was combined with the three personal valence items to form one factor while management support and efficacy constituted factors two and three. The three factor model combining appropriateness and valence was considered the most viable option because the content validity test and subsequent exploratory factor analysis suggested that participants had difficulty differentiating between personal and organizational benefits associated with the change, implying that two distinct dimensions may not exist.

The results indicated that the 25 items could not be adequately represented well by a single factor (NFI = 0.72; NNFI = 0.73; CFI = 0.72, RMSEA = 0.19). In contrast, the four factor model representing the four readiness for change factors that emerged in the exploratory analysis appeared to fit well (NFI = 0.96; NNFI = 0.96; CFI = 0.98, RMSEA = 0.08). That is, the values reported for the NFI, NNFI, and CFI exceeded 0.9 which is the typical cut-off score for these indices (where larger values indicate better fit) while the value of the RMSEA was 0.08 which is the typical cut-off value for this index (where smaller values indicate better fit) suggesting the four factor model was more suitable.

The fit of the hypothesized model was compared to other less complex models where the data were represented by two and three factors. In each of these cases, the values reported for the NFI, NNFI, and CFI exceeded the 0.9 cut-off score. However, the RMSEA value for each of these models was greater than the 0.08 cut-off value. To compare these models more directly and determine the extent to which the four factor model showed an improvement in fit over the other models, chi-square difference tests were performed. The proposed four factor model was compared to a three factor model. The three factor model had the following structure: (a) Factor 1 combined the three personal valence items with the ten appropriateness items; (b) Factor 2 consisted of the six management support items; and (c) Factor 3 consisted of the six efficacy items. A chi-square difference test indicated that the four factor model had significantly better fit than the three factor solution ($\Delta\chi^2 = 128.56$; df=4; $p < 0.01$).

In sum, the analyses suggested that the 25 items remaining after the confirmatory factor analysis constituted an acceptable version of the readiness factors. While the estimates of reliability were not as high in the second sample, the estimates of reliability were generally acceptable. Specifically, coefficient alphas were 0.80 for appropriateness, 0.79 for management support, 0.79 for change efficacy, and 0.65 for the personal valence score.

34.3.5.3 Convergent Validity Assessment

Personality and organizational factors were administered to this second sample as well. Means, standard deviations, correlations, and estimates of reliability are provided along the upper diagonal of Table 34.2. Consistent with previous findings, the results indicated that the readiness factors were correlated with each other (mean $r = 0.42$, $p < 0.05$). In addition, the correlations among the variables gave additional evidence of convergent validity. As expected, locus of control and general attitudes toward change were positively related to each of the readiness factors while rebelliousness was negatively related to each of the readiness factors. Additionally, perceptions of the communications climate and managements' ability were both positively related to the readiness for change factors.

34.4 Conclusion

This study was designed to construct a new instrument that measures readiness at an individual level because change activities are initiated and carried out by individuals within organizations. That is, even the most collective activities that take place within organizations are often an amalgamation of the activities of individual organizational members; therefore, organizations will accept or reject change through the actions of their members (Armenakis et al. 1993; Armenakisfield et al. 1999). To fulfill this objective, we took a series of steps that were designed to (a) specify the content domain of the readiness construct by integrating the strengths of the existing instruments, change theory, and manager experiences; (b) develop items to measure that domain; and (c) determine the extent to which items measure that domain. Furthermore, we wanted to test the instrument in a field setting.

The literature and published readiness for change instruments were coupled with qualitatively analyzed interviews and open-ended questionnaires from public- and private-sector managers. This analysis indicated that the most influential readiness factors, isolated empirically, were (a) discrepancy (i.e. the belief that a change was necessary), (b) efficacy (i.e. the belief that the change could be implemented), (c) organizational valence (i.e., the belief that the change would be organizationally beneficial), (d) management support (i.e., the belief that the organizational leaders were committed to the change), and (e) personal valence (i.e. the belief that the change would be personally beneficial). These five factors were formally defined; items were written to measure each; and two independent samples determined the extent to which the items reflected their intended constructs.

While the intended factor structure did not completely emerge (a point discussed later), the four scales that did emerge could be useful in an organizational setting. The factor structure was initially determined through exploratory methods in a public-sector organization and replicated with an independent sample, a private-sector organization, using confirmatory methods. The scales of appropriateness, management support, and change efficacy exceeded the minimum reliability estimate of 0.7. The personal valence scale was slightly less that this cut-off criterion (0.66 and 0.65 for the two organizations participating in the study). The scales also displayed convergent validity as evidenced by the correlations with personality and organizational variables (across two samples). Moreover, the measures distinguished between known groups (i.e. participants reported higher readiness than nonparticipants, as expected). Fourth, there was some evidence of predictive validity as demonstrated with the relationships between the readiness scales and three criteria (job satisfaction, affective commitment, and turnover intentions) measured seven months after the change. Finally, the readiness factors displayed incremental validity in predicting these outcomes after controlling for the organizational members personality (i.e. locus of control and general attitudes toward change) and perceptions of the context (i.e. communications climate and perceived management ability). When we tested for incremental validity, we found that the readiness for change factors did not explain a significant amount of variation in turnover intentions. One possible explanation for this may be that the mean age of the respondents was about 47. It is quite likely that this change may not have been unpopular enough for these respondents to anticipate terminating their employment.

In sum, there are several unique contributions made with this instrument and our process. First, we have followed a step-by-step process to develop our readiness measure. In doing this, we have provided some initial evidence of reliability and validity. Moreover, this provides a framework to evaluate other instruments that are currently available, facilitating wiser decisions as readiness is measured quantitatively (see Holt et al. 2007). Next, and potentially most important, the instrument is organizationally relevant and informs action. That is, this instrument taps specific attitudes that give insights into the messages that must be delivered to effectively initiate and implement change, thereby providing managers, organizational development consultants, and researchers an instrument that might best match their needs.

Despite the substantial support for the scales developed, there are several areas of concern. One basic concern was that our instrument was only tested in two organizations, both undergoing structural changes. Thus, the generalizability of the results across change types may be limited because it is reasonable to expect people to react to different types of changes differently. On the other hand, our instrument was built by capturing a range of participants' experiences where a wide range of educational (i.e. participants ranged from high school graduates to those with graduate degrees), functional (e.g. human resource management, engineering, flight operations, and education), and organizational backgrounds (i.e. public- and private-sector) were represented in our samples. This diversity offers some level of generalizability.

Another area of concern involves the extent to which the instrument is completely aligned with the readiness model presented. Specifically, the factors that were measured represented content (i.e. appropriateness), process (i.e. management support), and individual attributes (i.e. self-efficacy and personal valence) were represented. None the less, participants did not make the same distinctions between perceptions of discrepancy, organizational valence, and personal valence, omitting a context element (i.e. discrepancy). While the content validity assessments with the items indicated that these factors may be distinct, the results of the factor analysis suggested that there was considerable overlap in the constructs. So much so, the perceptions of discrepancy and organizational valence did not emerge as distinct factors in two separate samples. Therefore, there is a need for further refinements in these scales. At this point, we feel the refinements may be as simple as modifying the wording of the items. For example, one of the items intended to reflect personal valence was "I am concerned with the risks associated with this change". Recall that the participants in the content adequacy test I were almost equal in assigning it to personal valence and organizational valence. Consequently, we discarded the item. A closer examination of the item revealed that the item did not specify personal or organizational risks. Some may have interpreted it as organizational risks and some as personal risks. Furthermore, some of the items that were intended to refer to discrepancy were incorrectly worded. Specifically, the item "This change is clearly needed" is about a specific change. Discrepancy should refer to *a change* not *the change*. Organizational valence should include items that refer to the change having a benefit for the organization. Thus, the item "There are legitimate reasons for us to make this change" is more organizational valence than it is discrepancy. So, some refinements are needed in this instrument. After refinements are made, Hinkin's (1998) item development process can be fulfilled more completely. In particular, further testing of the instrument can be done to replicate the results and a confirmatory analysis of the scales conducted. Moreover, the predictive validity of the instrument warrants further attention. While our known-groups analysis was based on empirical findings (i.e. readiness would differ among groups of participants and nonparticipants), we cannot be assured that the groups did not differ along other dimensions besides participation, introducing selection bias into the findings.

On a more positive note, we feel the instrument can also be used to evaluate an implemented organizational change. It would be useful to change agents to know how the employees feel about proposed changes. Knowing whether or not the employees (a) felt the change was appropriate, (b) believed management supported the change, (c) felt capable to making the change successful, and (d) believed the change was personally beneficial would alert them to needed attention about the change. Periodic assessment of these sentiments may provide the necessary information to take whatever actions may be needed to make the change successful. Furthermore, such an instrument could be used in conjunction with other instruments that focus on measuring some aspect of change. We feel this instrument would be complementary to an instrument that assessed commitment to organizational change (cf. Herscovitch and Meyer 2002). If the commitment to change is not acceptable, the reasons may be in the dimensions assessed by this instrument.

In closing, this chapter has described the initial steps necessary to develop a valid and reliable instrument to assess readiness. Although the results that were reported here should be regarded as a preliminary step in developing an instrument to assess readiness, the results were encouraging. Despite the encouraging results, this effort sets the stage for a considerable research agenda. It has provided a framework to further explore the specific factors that influence readiness and a basis to build reliable and valid scales to measure those factors. Moreover, this can serve as framework to systematically assess facilitation strategies that can help leaders more effectively initiate and implement change.

References

Allen, N. J. and Meyer, J. P. (1990). The measurement and antecedents of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology,* 63, 1–18.

American Psychological Association. (1995). *Standards for Educational and Psychological Testing*. Washington DC.

Anderson, J. C. and Gerbing, D. W. (1991). Predicting the performance measures in a confirmatory factor analysis with a pretest assessment to their substantitive validity. *Journal of Applied Psychology,* 76(5), 732–740.

Armenakis, A. A., Harris, S. G., and Feild, H. S. (1999). Making change permanent: A model for institutionalizing change interventions. In *Research in Organizational Change and Development,* Vol. 12. New York: JAI Press, pp. 97–128.

Armenakis, A. A., Harris, S. G., and Mossholder, K. W. (1993). Creating readiness for organizational change. *Human Relations,* 46, 681–703.

Bolino, M. C., and Turnley, W. H. (1999). Measuring impression management in organizations: A scale development based on the Jones and Pittman taxonomy. *Organizational Research Methods,* 2(2), 187–206.

Boyle, G. J. (1991). Does item homogeneity indicate internal inconsistency or item redundancy in psychometric scales? *Personality and Individual Differences,* 12(3), 291–294.

Burke, W. W., Coruzzi, C. A., and Church, A. H. (Eds.). (1996). *The Organizational Survey as an Intervention for Change*. San Francisco: Jossey-Bass Publishers.

Cammann, C., Fichman, M., Jenkins, G. D., and Klesh, J. R. (1983). Assessing attitudes and perceptions of organizational members. In S. E. Seashore, E. E. Lawler, P. H. Mirvis, and C. Cammann (Eds.), *Assessing Organizational Change: A Guide to Methods, Measures, and Practices*. New York: John Wiley and Sons, pp. 71–138.

Coch, L. and French, J. R. P. (1948). Overcoming resistance to change. *Human Relations,* 1, 512–532.

Conway, J. M. and Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods,* 6, 147–168.

Cunningham, C. E., Woodward, C. A., Shannon, H. S., MacIntosh, J., Lendrum, B., Rosenbloom, D., and Brown, J. (2002). Readiness for organizational change: A longitudinal study of workplace, psychological, and behavioral correlates. *Journal of Occupational and Organizational Psychology,* 75, 377–392.

Ford, J. K., MacCallum, R. C., and Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology,* 39, 291–314.

Gopalakrishnan, S. and Damanpour, F. (2000). The impact of organizational context on innovation adoption in commercial banks. *IEEE Transactions on Engineering Management,* 47, 14–25.

Hersocovitch, L. and Meyer, J. P. (2002). Commitment to organizational change: Extension of a three-component model. *Journal of Applied Psychology,* 87, 474–487.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods,* 1, 104–121.

Holt, D. T., Armenakis, A. A., Harris, S. G., and Feild, H. S. (2007). Toward a comprehensive definition of readiness for change: A review of research and instrumentation. In W. A. Pasmore and R. W. Woodman (Eds.), *Research in Organizational Change and Development,* Vol. 16. Oxford, UK: Elsevier, pp. 289–336.

Hong, S. M. and Faedda, S. (1996). Refinement of the Hong psychological reactance scale. *Educational and Psychological Measurement,* 56, 173–182.

Isabella, L. A. (1990). Evolving interpretations as a change unfolds: How managers construe key organizational events. *Academy of Management Journal,* 33, 7–41.

Jacobson, E. H. (1957). *The effect of changing industrial methods and automation on personnel*. Paper presented at the Symposium on Preventive and Social Psychology, 15–17 April Washington, DC.

Jones, R. A., Jimmieson, N. L., and Griffiths, A. (2005). The impact of organizational culture and reshaping capabilities on change implementation success: The mediating role of readiness for change. *Journal of Management Studies,* 42, 893–899.

Kotter, J. P. (1995). Leading change: Why transformation efforts fail. *Harvard Business Review,* 73, 59–67.

Kotter, J. P. and Schlesinger, L. A. (1979). Choosing strategies for change. *Harvard Business Review,* 57, 106–114.

Lawrence, P. R. (1954). How to deal with resistance to change. *Harvard Business Review,* 32, 49–57.

Lewin, K. (1947). Frontiers in group dynamics: Concept, method and reality in social science; Social equilibria and social change. *Human Relations,* 1, 3–41.

Mayer, R. C. and Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology,* 84, 123–136.

McConnaughy, E. A., Prochaska, J. O., and Velicer, W. F. (1983). Stages of change in psychotherapy: Measurement and sample profiles. *Psychotherapy: Theory, Research, and Practice*, 20, 368–375.

Miller, K. I. and Monge, P. R. (1985). Social information and employee anxiety about organizational change. *Human Communication Research,* 11(3), 365–386.

Miller, V. D., Johnson, J. R., and Grau, J. (1994). Antecedents to willingness to participate in planned organizational change. *Journal of Applied Communications Research,* 22, 59–60.

Nunnally, J. C. (1978). *Psychometric Theory,* 2nd ed. New York: McGraw Hill Book Company.

Pearlin, L. I., Lieberman, M. A., Menaghan, E. G., and Mullan, J. T. (1981). The stress process. *Journal of Health and Social Behavior*, 22, 337–356.

Schriesheim, C. A. and Hinkin, T. R. (1990). Influence tactics used by subordinates: A theoretical and empirical analysis and refinement of the Kipnis Schmidt and Wilkinson subscales. *Journal of Applied Psychology,* 75, 246–257.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, G. C., and Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management,* 19, 385–417.

Sturman, M. C. and Short, J. C. (2000). Lump-Sum bonus satisfaction: Testing the construct validity of a new pay satisfaction dimension. *Personnel Psychology,* 53, 673–700.

Sutton, R. I. (1987). The process of organizational death: Disbanding and reconnecting. *Administrative Science Quarterly,* 32, 542–569.

Trumbo, D. A. (1961). Individual and group correlates of attitudes toward work-related change. *Journal of Applied Psychology,* 45, 338–344.

Van de Ven, Andrew H. and Marshall Scott Poole. (2005). Alternative approaches for studying organizational change. *Organization Studies*, 26(9), 1377–1404.

Wanberg, C. R. and Banas, J. T. (2000). Predictors and outcomes of openness to changes in a reorganizing workplace. *Journal of Applied Psychology,* 85, 132–142.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scale. *Journal of Personality and Social Psychology,* 54, 1063–1070.

Weeks, W. A., Roberts, J., Chonko, L. B., and Jones, E. (2004). Organizational readiness for change, individual fear of change, and sales manager performance: An empirical investigation. *Journal of Personal Selling and Sales Mangaement,* 24, 7–18.

Appendix A

Military Acronyms and Abbreviations

1NCD	1st Naval Construction Division
2-D	two-dimensional
2E	Role 2 enhanced
2LM	Role 2 light maneuver
3-D	three-dimensional
A	analog
A&P	administrative and personnel
A2C2	Army airspace command and control
A-3	Operations Directorate (COMAFFOR)
A-5	Plans Directorate (COMAFFOR)
AA	assessment agent; avenue of approach
AAA	antiaircraft artillery; arrival and assembly area; assign alternate area
AAAS	amphibious aviation assault ship
AABB	American Association of Blood Banks
AABWS	amphibious assault bulk water system
AAC	activity address code
AACG	arrival airfield control group
AADC	area air defense commander
AADP	area air defense plan
AA&E	arms, ammunition, and explosives
AAEC	aeromedical evacuation control team
AAFES	Army and Air Force Exchange Service
AAFIF	automated air facility information file
AAFS	amphibious assault fuel system
AAFSF	amphibious assault fuel supply facility
AAGS	Army air-ground system
AAI	air-to-air interface
AAM	air-to-air missile
AAMDC	US Army Air and Missile Defense Command
AAOE	arrival and assembly operations element
AAOG	arrival and assembly operations group
AAP	Allied administrative publication; assign alternate parent
AAR	after action report; after action review
AAST	aeromedical evacuation administrative support team
AAT	automatic analog test; aviation advisory team

AAU	analog applique unit
AAV	amphibious assault vehicle
AAW	antiair warfare
AB	airbase
ABCA	American, British, Canadian, Australian Armies Program
ABCS	Army Battle Command System
ABD	airbase defense
ABFC	advanced base functional component
ABFDS	aerial bulk fuel delivery system
ABFS	amphibious bulk fuel system
ABGD	air base ground defense
ABL	airborne laser
ABLTS	amphibious bulk liquid transfer system
ABM	antiballistic missile
ABN	airborne
ABNCP	Airborne Command Post
ABO	air base operability; blood typing system
A/C	aircraft
AC	active component; aircraft commander; alternating current
AC-130	Hercules
ACA	airlift clearance authority; airspace control authority; airspace coordination area
ACAA	automatic chemical agent alarm
ACAPS	area communications electronics capabilities
ACAT	aeromedical evacuation command augmentation team
ACB	amphibious construction battalion
ACC	Air Combat Command; air component commander; area coordination center
ACCE	air component coordination element
ACCON	acoustic condition
ACCS	air command and control system
ACCSA	Allied Communications and Computer Security Agency
ACDO	assistant command duty officer
ACE	airborne command element (USAF); air combat element (NATO); Allied Command Europe; aviation combat element Marine air-ground task force (MAGTF)
ACEOI	Automated Communications-Electronics Operating Instructions
ACF	air contingency force
ACI	assign call inhibit
ACIC	Army Counterintelligence Center
ACINT	acoustic intelligence
ACK	acknowledgement
ACL	access control list; allowable cabin load
ACLANT	Allied Command Atlantic
ACM	advanced conventional munitions; advanced cruise missile; air combat maneuver; air contingency Marine air-ground task force (MAGTF); airspace coordinating measure
ACMREQ	airspace control means request; airspace coordination measures request

ACN	assign commercial network
ACO	administrative contracting officer; airspace control order
ACOC	area communications operations center
ACOCC	air combat operations command center
ACOS	assistant chief of staff
ACP	access control point; air commander's pointer; airspace control plan; Allied Communications Publication; Assign common pool
ACR	armored cavalry regiment (Army); assign channel reassignment
ACS	agile combat support; air-capable ship; airspace control system; auxiliary crane ship
ACSA	acquisition and cross-servicing agreement; Allied Communications Security Agency
AC/S, C4I	Assistant Chief of Staff, Command, Control, Communications, Computers, and Intelligence (USMC)
ACT	activity; Allied Command Transformation
ACU	assault craft unit
ACV	aircraft cockpit video; air cushion vehicle; armored combat vehicle
ACW	advanced conventional weapons
A/D	analog-to-digital
AD	active duty; advanced deployability; air defense; automatic distribution; priority add-on
ADA	aerial damage assessment; air defense artillery
A/DACG	arrival/departure airfield control group
ADAFCO	air defense artillery fire control officer
ADAL	authorized dental allowance list
ADAM/BAE	air defense airspace management/brigade aviation element
ADAMS	Allied Deployment and Movement System
ADANS	Air Mobility Command Deployment Analysis System
ADC	air defense commander; area damage control
ADCAP	advanced capability
A/DCG	arrival/departure control group
ADCI/MS	Associate Director of Central Intelligence for Military Support
ADCON	administrative control
ADD	assign on-line diagnostic
ADDO	Assistant Deputy Director for Operations
ADDO(MS)	Assistant Deputy Director for Operations/Military Support
ADE	air defense emergency; assign digit editing
ADF	automatic direction finding
ADIZ	air defense identification zone
ADKC/RCU	Automatic Key Distribution Center/Rekeying Control Unit
ADL	advanced distributed learning; armistice demarcation line; assign XX (SL) routing
ADMIN	administration
ADN	Allied Command Europe desired ground zero number
ADNET	anti-drug network
ADOC	air defense operations center
ADP	air defense plan; automated data processing
ADPE	automated data processing equipment

ADPS	automatic data processing system
ADR	accident data recorder; aircraft damage repair; armament delivery recording
ADRA	Adventist Development and Relief Agency
ADS	air defense section; air defense sector; amphibian discharge site
ADSIA	Allied Data Systems Interoperability Agency
ADSW	active duty for special work
ADT	active duty for training; assign digital transmission group; automatic digital tester
ADUSD(TP)	Assistant Deputy Under Secretary of Defense, Transportation Policy
ADVON	advanced echelon
ADW	air defense warnings
ADWC	air defense warning condition
ADZ	amphibious defense zone
AE	aeromedical evacuation; assault echelon; attenuation equalizer
AEC	aeromedical evacuation crew
AECA	Arms Export Control Act
AECC	aeromedical evacuation coordination center
AECM	aeromedical evacuation crew member
AECS	aeromedical evacuation command squadron
AECT	aeromedical evacuation control team
AEF	air and space expeditionary force
AEG	air expeditionary group
AELT	aeromedical evacuation liaison team
AEOS	aeromedical evacuation operations squadron
AEOT	aeromedical evacuation operations team
AEPS	aircrew escape propulsion system
AEPST	aeromedical evacuation plans and strategy team
AES	aeromedical evacuation squadron; aeromedical evacuation system
AESC	aeromedical evacuation support cell
AETC	Air Education and Training Command
AETF	air and space expeditionary task force
A/ETF	automated/electronic target folder
AEU	assign essential user bypass
AEW	air and space expeditionary wing; airborne early warning
AEW&C	airborne early warning and control A-5
AF	amphibious force
AFAARS	Air Force After Action Reporting System
AFAOC	Air Force air and space operations center
AFARN	Air Force air request net
AFATDS	Advanced Field Artillery Tactical Data System
AFB	Air Force base
AFC	area frequency coordinator; automatic frequency control
AFCA	Air Force Communications Agency
AFCAP	Air Force contract augmentation program; Armed Forces contract augmentation program
AFCC	Air Force Component Commander
AFCCC	Air Force Combat Climatology Center
AFCEE	Air Force Center for Environmental Excellence
AFCENT	Allied Forces Central Europe (NATO)

AFCERT	Air Force computer emergency response team
AFCESA	Air Force Civil Engineering Support Agency
AFCS	automatic flight control system
AFD	assign fixed directory
AFDC	Air Force Doctrine Center
AFDD	Air Force doctrine document
AFDIGS	Air Force digital graphics system
AFDIL	Armed Forces DNA Identification Laboratory
AFDIS	Air Force Weather Agency Dial In Subsystem
AF/DP	Deputy Chief of Staff for Personnel, United States Air Force
AFE	Armed Forces Entertainment
AFEES	Armed Forces Examining and Entrance Station
AFFIS	Air Facilities File Information System
AFFMA	Air Force Frequency Management Agency
AFFOR	Air Force forces
AFH	Air Force handbook
AFI	Air Force instruction
AFID	anti-fratricide identification device
AF/IL	Deputy Chief of Staff for Installations and Logistics, USAF
AFIT	Air Force Institute of Technology
AFIP	Armed Forces Institute of Pathology
AFIS	American Forces Information Service
AFIRB	Armed Forces Identification Review Board
AFIWC	Air Force Information Warfare Center
AFJI	Air Force joint instruction
AFJMAN	Air Force Joint Manual
AFLC	Air Force Logistics Command
AFLE	Air Force liaison element
AFLNO	Air Force liaison officer
AFMAN	Air Force manual
AFMC	Air Force Materiel Command A-6
AFMD	Air Force Mission Directive
AFME	Armed Forces Medical Examiner
AFMES	Armed Forces Medical Examiner System
AFMIC	Armed Forces Medical Intelligence Center
AFMLO	Air Force Medical Logistics Office
AFMS	Air Force Medical Service
AFNORTH	Air Force North; Allied Forces Northern Europe (NATO)
AFNW	Allied Forces North West Europe (NATO)
AFNSEP	Air Force National Security and Emergency Preparedness Agency
AFOE	assault follow-on echelon
AFOSI	Air Force Office of Special Investigations
AFPAM	Air Force pamphlet
AFPC	Air Force Personnel Center
AFPD	Air Force policy directive
AFPEO	Armed Forces Professional Entertainment Overseas
AFR	Air Force Reserve; assign frequency for network reporting
AFRC	Air Force Reserve Command; Armed Forces Recreation Center
AFRCC	Air Force rescue coordination center
AFRRI	Armed Forces Radiobiology Research Institute

AFRTS	Armed Forces Radio and Television Service
AFS	aeronautical fixed service
AFSATCOM	Air Force satellite communications (system)
AFSC	Armed Forces Staff College; United States Air Force specialty code
AFSOB	Air Force special operations base
AFSOC	Air Force Special Operations Command; Air Force special operations component
AFSOCC	Air Force special operations control center
AFSOD	Air Force special operations detachment
AFSOE	Air Force special operations element
AFSOF	Air Force special operations forces
AFSOUTH	Allied Forces, South (NATO)
AFSPACE	United States Space Command Air Force
AFSPC	Air Force Space Command
AFSPOC	Air Force Space Operations Center
AFTAC	Air Force Technical Applications Center
AFTH	Air Force Theater Hospital
AFTN	Aeronautical Fixed Telecommunications Network
AFTO	Air Force technical order
AFTTP	Air Force tactics, techniques, and procedures; Air Force technical training publication
AFTTP(I)	Air Force tactics, techniques, and procedures (instruction)
AFW	Air Force Weather
AFWA	Air Force Weather Agency
AFWCF	Air Force working capital fund
AFWIN	Air Force Weather Information Network
AF/XO	Deputy Chief of Staff for Plans and Operations, United States Air Force
AF/XOI	Air Force Director of Intelligence, Surveillance, and Reconnaissance
AF/XOO	Director of Operations, United States Air Force
A/G	air to ground
AG	adjutant general (Army)
AGARD	Advisory Group for Aerospace Research and Development
AGCCS	Army Global Command and Control System
AGE	aerospace ground equipment
AGI	advanced geospatial intelligence
AGIL	airborne general illumination lightself
AGL	above ground level
AGM-28A	Hound Dog
AGM-65	Maverick
AGM-69	short range attack missile
AGR	Active Guard and Reserve
AGS	aviation ground support
AHA	alert holding area
AHD	antihandling device
AI	airborne interceptor; air interdiction; area of interest
AIA	Air Intelligence Agency
AIASA	annual integrated assessment for security assistance
AIC	air intercept controller; assign individual compressed dial; Atlantic Intelligence Command

AICF/USA	Action Internationale Contre La Faim (International Action Against Hunger)
AIDS	acquired immune deficiency syndrome
AIF	automated installation intelligence file
AIFA	AAFES Imprest Fund Activity
AIG	addressee indicator group
AIIRS	automated intelligence information reporting system
AIK	assistance in kind
AIM	Airman's Information Manual
AIM-7	Sparrow
AIM-9	Sidewinder
AIM-54	A Phoenix
AIMD	aircraft intermediate maintenance department
AIQC	antiterrorism instructor qualification course
AIRBAT	Airborne Intelligence, Surveillance, and Reconnaissance Requirements-Based Allocation Tool
AIRCENT	Allied Air Forces Central Europe (NATO)
AIRES	advanced imagery requirements exploitation system
AIREVAC	air evacuation (confirmation)
AIREVAC	air evacuation (request)
AIREVAC	air evacuation (response)
AIRNW	Allied Air Forces North West Europe (NATO)
AIRREQ	air request (reconnaissance)
AIRSOUTH	Allied Air Forces Southern Europe (NATO)
AIRSUPREQ	air support request
AIS	automated information system
AIT	aeromedical isolation team; automatic identification technology
AIU	Automatic Digital Network Interface Unit
AJ	anti-jam
AJBPO	area joint blood program office
AJCC	alternate joint communications center
AJ/CM	anti-jam control modem
AJF	allied joint force
AJFP	adaptive joint force packaging
AJMRO	area joint medical regulating office
AJNPE	airborne joint nuclear planning element
AJP	Allied joint publication
AK	commercial cargo ship
AKNLDG	acknowledge message
ALCC	airlift control center; airlift coordination cell
ALCE	airlift control element
ALCF	airlift control flight
ALCG	analog line conditioning group
ALCM	air launched cruise missile
ALCOM	United States Alaskan Command
ALCON	all concerned
ALCS	airlift control squadron
ALCT	airlift control team
ALD	accounting line designator; airborne laser designator; available-to-load date

ALE	airlift liaison element
ALERFA	alert phase (ICAO)
ALERT	attack and launch early reporting to theater
ALERTORD	alert order
ALLOREQ	air allocation request
ALLTV	all light level television
ALMSNSCD	airlift mission schedule
ALNOT	alert notice; search and rescue alert notice
ALO	air liaison officer
ALOC	air line of communications
ALORD	alert launch order
ALP	Allied Logistic Publication
ALSA	Air Land Sea Application (Center)
ALSS	advanced logistic support site
ALTRV	altitude reservation
ALTTSC	alternate Tomahawk strike coordinator
A/M	approach and moor
AM	amplitude modulation
AMAL	authorized medical allowance list
AMB	air mobility branch; ambassador
AMBUS	ambulance bus
AMC	airborne mission coordinator; Air Mobility Command; Army Materiel Command: midpoint compromise search area
AMCC	allied movement coordination center; alternate military command center
AMCIT	American citizen
AMCM	airborne mine countermeasures
AMCT	air mobility control team
AMD	air and missile defense; air mobility division
AME	air mobility element; antenna mounted electronics
AMEDD	Army Medical Department
AMEDDCS	US Army Medical Department Center and School
AMEMB	American Embassy
AMF(L)	ACE Mobile Force (Land) (NATO)
AMH	automated message handler
AMIO	alien migrant interdiction operations
AMLO	air mobility liaison officer
AMMO	ammunition
AMOC	Air Marine Operations Center
AMOCC	air mobility operations control center
AMOG	air mobility operations group
AMOPES	Army Mobilization and Operations Planning and Execution System
AMOPS	Army mobilization and operations planning system; Army mobilization operations system
AMOS	air mobility operations squadron
AMOSS	Air and Marine Operations Surveillance System
AMP	amplifier; analysis of mobility platform
AMPE	automated message processing exchange
AMPN	amplification
AMPSSO	Automated Message Processing System Security Office (or Officer)
AMRAAM	advanced medium-range air-to-air missile

AMS	aerial measuring system; air mobility squadron; Army management structure; Asset Management System A-10
AMSS	air mobility support squadron
AMT	aerial mail terminal
AMVER	automated mutual-assistance vessel rescue system
AMW	air mobility wing; amphibious warfare
AMX	air mobility express
AN	alphanumeric; analog nonsecure
ANCA	Allied Naval Communications Agency
ANDVT	advanced narrowband digital voice terminal
ANG	Air National Guard
ANGUS	Air National Guard of the United States
A/NM	administrative/network management
ANMCC	Alternate National Military Command Center
ANN	assign NNX routing
ANR	Alaskan North American Aerospace Defense Command Region
ANSI	American National Standards Institute
ANX	assign NNXX routing
ANY	assign NYX routing
ANZUS	Australia–New Zealand–United States Treaty
AO	action officer; administration officer; air officer; area of operations; aviation ordnance person
AO&M	administration, operation, and maintenance
AOA	amphibious objective area
AOB	advanced operations base; aviation operations branch
AOC	air operations center; Army operations center
AOCC	air operations control center
AOC-E	Aviation Operations Center-East (USCS)
AOCU	analog orderwire control unit
AOC-W	Aviation Operations Center-West (USCS)
AOD	on-line diagnostic
AOF	azimuth of fire
AOI	area of interest
AOL	area of limitation
AOP	air operations plan; area of probability
AOR	area of responsibility
AOS	area of separation
AOSS	aviation ordnance safety supervisor
AOTR	Aviation Operational Threat Response
AP	allied publication; antipersonnel; average power
APA	Army pre-positioned afloat
APAN	Asia Pacific Network
APC	aerial port commander; armored personnel carrier; assign preprogrammed conference list
APCC	aerial port control center; alternate processing and correlation center
APES	Automated Patient Evacuation System
APF	afloat pre-positioning force
APG	aimpoint graphic
APHIS	Animal and Plant Health Inspection Service

APIC	allied press information center
APL	antipersonnel land
APO	afloat pre-positioning operations; Army Post Office
APOD	aerial port of debarkation
APOE	aerial port of embarkation
APORT	aerial port
APP	allied procedural publication
APPS	analytical photogrammetric positioning system
APR	assign primary zone routing
APS	aerial port squadron; afloat pre-positioning ship; Army pre-positioned stocks
APS-3	afloat pre-positioning stocks
APU	auxiliary power unit
AR	air refueling; Army regulation; Army reserve
ARB	alternate recovery base; assign receive bypass lists
ARBS	angle rate bombing system
ARC	air Reserve Components; American Red Cross
ARCENT	United States Army Central Command
ARCP	air refueling control point
ARCT	air refueling control team; air refueling control time
ARDF	automatic radio direction finding
AREC	air resource element coordinator
ARFOR	Army forces
ARG	amphibious ready group
ARGO	automatic ranging grid overlay
ARINC	Aeronautical Radio Incorporated
ARIP	air refueling initiation point
ARL-M	airborne reconnaissance low-multifunction
ARM	antiradiation missile
ARNG	Army National Guard
ARNGUS	Army National Guard of the United States
ARNORTH	US Army North
ARP	air refueling point
ARPERCEN	US Army Reserve Personnel Center
ARQ	automatic request-repeat
ARRC	Allied Command Europe Rapid Reaction Corps (NATO)
ARRDATE	arrival date
ARS	acute radiation syndrome; air rescue service
ARSOA	Army special operations aviation
ARSOC	Army special operations component
ARSOF	Army special operations forces
ARSOTF	Army special operations task force
ARSPACE	Army Space Command
ARSPOC	Army space operations center
ART	air reserve technician
ARTCC	air route traffic control center
ARTS III	Automated Radar Tracking System
ARTYMET	artillery meteorological
AS	analog secure; aviation ship
A/S	anti-spoofing

ASA	automatic spectrum analyzer
ASAP	as soon as possible
ASARS	Advanced Synthetic Aperture Radar System
ASAS	All Source Analysis System
ASAT	antisatellite weapon
ASB	naval advanced support base
ASBP	Armed Services Blood Program
ASBPO	Armed Services Blood Program Office
ASC	acting service chief; Aeronautical Systems Center; Air Systems Command; assign switch classmark; Automatic Digital Network switching center
ASCC	Air Standardization Coordinating Committee; Army Service component command; Army Service component commander
ASCIET	all Services combat identification evaluation team
ASCII	American Standard Code for Information Interchange
ASCS	air support control section
ASD(A&L)	Assistant Secretary of Defense (Acquisition and Logistics)
ASD(C)	Assistant Secretary of Defense (Comptroller)
ASD(C3I)	Assistant Secretary of Defense (Command, Control, Communications, and Intelligence)
ASD(FM&P)	Assistant Secretary of Defense (Force Management and Personnel)
ASD(FMP)	Assistant Secretary of Defense (Force Management Policy)
ASD(HA)	Assistant Secretary of Defense (Health Affairs)
ASD(HD)	Assistant Secretary of Defense (Homeland Defense)
ASD	Assistant Secretary of Defense (for HD & ASA: Homeland Defense and Americas' Security Affairs)
ASDI	analog simple data interface
ASDIA	All-Source Document Index
ASD(ISA)	Assistant Secretary of Defense (International Security Affairs)
ASD(ISP)	Assistant Secretary of Defense (International Security Policy)
ASD(LA)	Assistant Secretary of Defense (Legislative Affairs)
ASD(NII)	Assistant Secretary of Defense (Networks and Information Integration)
ASD(P&L)	Assistant Secretary of Defense (Production and Logistics)
ASD(PA)	Assistant Secretary of Defense (Public Affairs)
ASD(PA&E)	Assistant Secretary of Defense (Program Analysis and Evaluation)
ASD(RA)	Assistant Secretary of Defense (Reserve Affairs)
ASD(RSA)	Assistant Secretary of Defense (Regional Security Affairs)
ASD(S&R)	Assistant Secretary of Defense (Strategy and Requirements)
ASD	Assistant Secretary of Defense (for SO/LIC: Special Operations and Low-Intensity Conflict)
ASD	Assistant Secretary of Defense (for SO/LIC & IC: Special Operations and Low-Intensity Conflict and Interdependent Capabilities)
ASE	aircraft survivability equipment; automated stabilization equipment
ASF	aeromedical staging facility
ASG	area support group
ASH	Assistant Administrator for Security and Hazardous Materials
ASI	assign and display switch initialization
ASIC	Air and Space Interoperability Council
ASIF	Airlift Support Industrial Fund

ASL	allowable supply list; archipelagic sea lane; assign switch locator (SL) routing; authorized stockage list (Army)
ASM	air-to-surface missile; armored scout mission; automated scheduling message
ASMD	antiship missile defense
ASO	advanced special operations; air support operations
ASOC	air support operations center
ASOFDTG	as of date/time group
ASPA	American Service-Members' Protection Act
ASPP	acquisition systems protection program
ASPPO	Armed Service Production Planning Office
ASR	available supply rate
ASSETREP	transportation assets report
AST	assign secondary traffic channels
ASTS	aeromedical staging squadron
ASW	antisubmarine warfare; average surface wind
ASWBPL	Armed Services Whole Blood Processing Laboratories
ASWC	antisubmarine warfare commander
AT	annual training; antitank; antiterrorism
At	total attainable search area
ATA	Airlift Tanker Association; airport traffic area
ATAC	antiterrorism alert center (Navy)
ATACC	advanced tactical air command center
ATACMS	Army Tactical Missile System
ATACO	air tactical actions control officer
ATACS	Army Tactical Communications System
ATAF	Allied Tactical Air Force (NATO)
ATBM	antitactical ballistic missile
ATC	air target chart; Air Threat Conference; air traffic control; air transportable clinic (USAF)
ATCA	Allied Tactical Communications Agency
ATCAA	air traffic control assigned airspace
ATCALS	air traffic control and landing system
ATCC	air traffic control center; Antiterrorism Coordinating Committee
ATCC-SSG	Antiterrorism Coordinating Committee-Senior Steering Group
ATCRBS	Air Traffic Control Radar Beacon System
ATCS	air traffic control section
ATDL1	Army tactical data link 1
ATDLS	Advanced Tactical Data Link System
ATDM	adaptive time division multiplexer
ATDS	airborne tactical data system
ATEP	Antiterrorism Enterprise Portal
ATF	Advanced Targeting FLIR; amphibious task force; Bureau of Alcohol, Tobacco and Firearms (TREAS)
AT/FP	antiterrorism/force protection
ATG	amphibious task group; assign trunk group cluster
ATGM	antitank guided missile; antitank guided munition
ATH	air transportable hospital; assign thresholds
ATHS	Airborne Target Handover System
ATM	advanced trauma management; air target material; assign traffic metering

ATMCT	air terminal movement control team
ATMP	Air Target Materials Program
ATN	assign thresholds
ATO	air tasking order; antiterrorism officer
ATOC	air tactical operations center; air terminal operations center
ATOCONF	air tasking order/confirmation
ATP	allied tactical publication
ATR	attrition reserve
ATS	air traffic service; assign terminal service
ATSD(AE)	Assistant to the Secretary of Defense (Atomic Energy)
ATSD(IO)	Assistant to the Secretary of Defense (Intelligence Oversight)
ATSD(NCB)	Assistant to the Secretary of Defense for Nuclear and Chemical and Biological Defense Programs
ATT	assign terminal type
ATTU	air transportable treatment unit
ATWG	antiterrorism working group
AUEL	automated unit equipment list
AUF	airborne use of force
AUG	application user group
AUIC	active duty unit identification code
AUTODIN	Automatic Digital Network
AUX	auxiliary
AV	air vehicle; asset visibility
AV-8	Harrier
AVDTG	analog via digital trunk group
AVGAS	aviation gasoline
AVIM	aviation intermediate maintenance
AVL	assign variable location
AVOU	analog voice orderwire unit
AVOW	analog voice orderwire
AVS	audiovisual squadron
AVUM	aviation unit maintenance
AV/VI	audiovisual/visual information
AW	acoustic warfare; air warfare
AWACS	Airborne Warning and Control System
AWADS	adverse weather aerial delivery system
AWC	air warfare commander
AWCAP	airborne weapons corrective action program
AWCCM	acoustic warfare counter-countermeasures
AWCM	acoustic warfare countermeasures
AWDS	automated weather distribution system
AWN	Automated Weather Network
AWOL	absent without leave
AWS	Air Weather Service
AWSE	armament weapons support equipment
AWSIM	air warfare simulation model
AWSR	Air Weather Service regulation
AXP	ambulance exchange point
AXX	assign XXX routing
AZR	assign zone restriction lists

B	cross-over barrier pattern
B-52	Stratofortress
B&A	boat and aircraft
BAF	backup alert force
BAG	baggage
BAH	basic allowance for housing
BAI	backup aircraft inventory; battlefield air interdiction
BAS	basic allowance for subsistence; battalion aid station
BATF	Bureau of Alcohol, Tobacco, and Firearms
B/B	baseband
BB	breakbulk
BBL	barrel (42 US gallons)
BC	bottom current
BCA	border crossing authority
BCAT	beddown capability assessment tool
BCD	battlefield coordination detachment
BCI	bit count integrity
BCN	beacon
BCOC	base cluster operations center
BCR	baseline change request
BCT	brigade combat team
BCTP	battle command training program
BCU	beach clearance unit
BDA	battle damage assessment
BDAREP	battle damage assessment report
BDC	blood donor center
BDE	brigade
BDL	beach discharge lighter
BDOC	base defense operations center
BDR	battle damage repair
BDRP	Biological Defense Research Program
BDZ	base defense zone
BE	basic encyclopedia
BEAR	base expeditionary airfield resources
BEE	bioenvironmental engineering officer
BEN	base encyclopedia number
BE number	basic encyclopedia number
BER	bit error ratio
BES	budget estimate submission
BfV	*Bundesamt für Verfassungsschutz* (federal office for defending the Constitution)
BGC	boat group commander
BHR	Bureau of Humanitarian Response
BI	battlefield injury
BIA	Bureau of Indian Affairs
BIAS	Battlefield Illumination Assistance System
BIDDS	Base Information Digital Distribution System
BIDE	basic identity data element
BIFC	Boise Interagency Fire Center
BIH	International Time Bureau (Bureau International d'l'Heure)

BII	base information infrastructure
BINM	Bureau of International Narcotics Matters
BIO	biological; Bureau of International Organizations
BISS	base installation security system
BIT	built-in test
BITE	built-in test equipment
BIU	beach interface unit
BKA	*Bundeskriminalamt* (federal criminal office)
BL	biocontainment level
BLCP	beach lighterage control point
BLDREP	blood report
BLDSHIP	blood shipment (report)
BLM	Bureau of Land Management
BLOS	beyond line of sight
BLS	beach landing site
BLT	battalion landing team
BM	ballistic missile; battle management; beachmaster
BMC4I	Battle Management Command, Control, Communications, Computers, and Intelligence
BMCT	begin morning civil twilight
BMD	ballistic missile defense
BMDO	Ballistic Missile Defense Organization
BMET	biomedical equipment technician
BMEWS	ballistic missile early warning system
BMNT	begin morning nautical twilight
BMU	beachmaster unit
BN	battalion
BND	*Bundesnachrichtendienst* (federal intelligence service)
BOA	basic ordering agreement
BOC	base operations center
BOCCA	Bureau of Coordination of Civil Aircraft (NATO)
BOG	beach operations group
BOH	bottom of hill
BORFIC	Border Patrol Field Intelligence Center
BOS	base operating support; battlefield operating system
BOSG	base operations support group
BOSS	base operating support service
BP	battle position; block parity
BPA	blanket purchase agreement
BPD	blood products depot
BPG	beach party group
BPI	bits per inch
BPO	blood program office
BPPBS	bi-annual planning, programming, and budget system
bps	bits per second
BPSK	biphase shift keying
BPT	beach party team
BPWRR	bulk petroleum war reserve requirement
BPWRS	bulk petroleum war reserve stocks
BR	budget review

BRAC	base realignment and closure
BRACE	Base Resource and Capability Estimator
BRC	base recovery course
BS	battle staff; broadcast source
BSA	beach support area; brigade support area
BSC	black station clock
BSC ro	black station clock receive out
BSCT	behavioral science consultation team
BSD	blood supply detachment
BSI	base support installation
BSP	base support plan
BSSG	brigade service support group
BSU	blood supply unit
BT	bathythermograph
BTB	believed-to-be
BTC	blood transshipment center
BTG	basic target graphic
BTOC	battalion tactical operations center
BTS	Border and Transportation Security (DHS)
BTU	beach termination unit
BULK	bulk cargo
BUMED	Bureau of Medicine and Surgery (instruction)
BVR	beyond visual range
BW	bandwidth; biological warfare; biological weapon
BWC	Biological Weapons Convention
BZ	buffer zone
C	centigrade; clock; compromise band; coverage factor; creeping line pattern
C&A	certification and accreditation
C&E	communications and electronics
C&LAT	cargo and loading analysis table
C2	command and control
C2-attack	an offensive form of command and control warfare
C2E	command and control element
C2IP	Command and Control Initiatives Program
C2IPS	Command and Control Information Processing System
C2P	command and control protection
C2-protect	a defensive form of command and control warfare
C2S	command and control support
C-2X	coalition Intelligence Directorate counterintelligence and human intelligence staff element
C3	command, control, and communications
C3AG	Command, Control, and Communications Advisory Group
C3CM	command, control, and communications countermeasures
C3I	command, control, communications, and intelligence
C3IC	coalition coordination, communications, and integration center
C3SMP	Command, Control, and Communications Systems Master Plan

C4CM	command, control, communications, and computer countermeasures
C4I	command, control, communications, computers, and intelligence
C4IFTW	command, control, communications, computers, and intelligence for the Warrior
C4ISR	command, control, communications, computers, intelligence, surveillance, and reconnaissance
C4S	command, control, communications, and computer systems
C4	systems command, control, communications, and computer systems
C-5	Galaxy
C-17	Globemaster III
C-21	Learjet
C-27	Spartan
C-130	Hercules
C-141	Starlifter
CA	civil administration; civil affairs; combat assessment
C/A	course acquisition
CAA	civil air augmentation; combat aviation advisors; command arrangement agreement
CAB	combat aviation brigade
CAC	common access card; current actions center
CACOM	Civil Affairs command
CACTIS	community automated intelligence system
CAD	Canadian Air Division; cartridge actuated device; collective address designator
CADRS	concern and deficiency reporting system
CADS	containerized ammunition distribution system
CAE	command assessment element
CAF	Canadian Air Force; combat air forces; commander, airborne/air assault force
CAFMS	computer-assisted force management system
CAG	carrier air group; civil affairs group; collective address group
CAIMS	conventional ammunition integrated management system
CAINS	carrier aircraft inertial navigation system
CAL	caliber; critical asset list
CALA	Community Airborne Library Architecture
CALCM	conventional air-launched cruise missile
CALICS	communication, authentication, location, intentions, condition, and situation
CALMS	computer-aided load manifesting system
CAM	chemical agent monitor; crisis action module
CAMPS	Consolidated Air Mobility Planning System
CAMT	countering air and missile threats
CANA	convulsant antidote for nerve agent
CANR	Canadian North American Aerospace Defense Command Region
CANUS	Canada–United States
CAO	chief administrative officer; civil affairs operations; counterair operation
CAOC	combat air operations center; combined air operations center
CAO SOP	standing operating procedures for coordination of atomic operations

CAP	Civil Air Patrol; civil augmentation program; combat air patrol; configuration and alarm panel; Consolidated Appeals Process (UN); crisis action planning
CAR	Chief of the Army Reserve
CARDA	continental United States airborne reconnaissance for damage assessment; continental United States area reconnaissance for damage assessment
CARE	Cooperative for Assistance and Relief Everywhere (CAREUSA)
CARIBROC	Caribbean Regional Operations Center
CARP	computed air release point; contingency alternate route plan
CARS	combat arms regimental system
CARVER	criticality, accessibility, recuperability, vulnerability, effect, and recognizability
CAS	casualty; civil aviation security; close air support
CASEVAC	casualty evacuation
CASF	contingency aeromedical staging facility
CASP	computer-aided search planning
CASPER	contact area summary position report
CASREP	casualty report
CASREQ	close air support request
CAT	category; crisis action team
CATCC	carrier air traffic control center
CATF	commander, amphibious task force
CAU	crypto ancillary unit; cryptographic auxiliary unit
CAVU	ceiling and visibility unlimited
CAW	carrier air wing
CAW/ESS	crisis action weather and environmental support system
CAX	computer-assisted exercise
C-B	chemical-biological
CB	chemical-biological; construction battalion (SEABEES)
CBBLS	hundreds of barrels
CBD	chemical, biological defense
CBFS	cesium beam frequency standard
CBIRF	chemical-biological incident response force
CBLTU	common battery line terminal unit
CBMR	capabilities-based munitions requirements
CBMU	construction battalion maintenance unit
CBP	capabilities-based planning; Customs and Border Protection
CBPO	Consolidated Base Personnel Office
CBPS	chemical biological protective shelter
CBR	chemical, biological, and radiological
CBRN	Caribbean Basin Radar Network; chemical, biological, radiological, and nuclear
CBRNE	chemical, biological, radiological, nuclear, and high-yield explosives
CBRT	chemical–biological response team
CBS	common battery signaling
CBT	common battery terminal
CbT	combating terrorism
CbT-RIF	Combating Terrorism Readiness Initiatives Fund
CBTZ	combat zone

CBU	cluster bomb unit; conference bridge unit; construction battalion unit
CBW	chemical and biological warfare
C/C	cabin cruiser; cast off and clear
CC	command center; component command (NATO)
CC&D	camouflage, concealment, and deception
CCA	carrier-controlled approach; central contracting authority; circuit card assembly; container control activity; contamination control area; contingency capabilities assessment; contract construction agent (DOD)
CCAP	combatant command AFRTS planner
CCAS	contingency contract administration services
CCAS-C	contingency contract administration services commander
CCATT	critical care air transport team
CCB	Community Counterterrorism Board; Configuration Control Board
CCC	communication, cooperation, coordination (Badiru's project management model) coalition coordination cell; coalition coordination center; Combined Command Center (USSPACECOM); crisis coordination center; critical control circuit; crosscultural communications course
CCD	camouflage, concealment, and deception
CCDR	combatant commander
CCE	container control element; continuing criminal enterprise
CCEB	Combined Communications-Electronics Board
CCF	collection coordination facility
CCG	crisis coordination group
CCGD	commander, Coast Guard district
CCIB	command center integration branch
CCIF	Combatant Commander's Initiative Fund
CCIP	continuously computed impact point
CCIR	commander's critical information requirement; International Radio Consultative Committee
CCIS	common channel interswitch signaling
CCITT	International Telegraph and Telephone Consultative Committee
CCIU	CEF control interface unit
CCJTF	commander, combined joint task force
CCL	communications/computer link
CCLI	computer control list item
CCO	central control officer; combat cargo officer; command and control office; complex contingency operation
CCOI	critical contact of interest
CCP	casualty collection point; consolidated cryptologic program; consolidation and containerization point
CCPDS	command center processing and display system
CCRD	combatant commander's required delivery date
C-CS	communication and computer systems
CCS	central control ship; container control site
CCSA	containership cargo stowage adapter
CCSD	command communications service designator; control communications service designator
CCT	collaborative contingency targeting; combat control team
CCTI	Chairman of the Joint Chiefs of Staff commended training issue

CCTV	closed circuit television
CCW	Convention on Conventional Weapons (1980 United Nations); continuous carrier wave
CD	channel designator; compact disc; counterdrug
C-day	unnamed day on which a deployment operation begins
CDC	Centers for Disease Control and Prevention
CDE	collateral damage estimation
CDF	combined distribution frame; contractors deploying with the force
CDI	cargo disposition instructions; conditioned diphase C di conditioned diphase
CDIP	combined defense improvement project
CDIPO	counterdrug intelligence preparation for operations
CDLMS	common data link management system
CDM	cable driver modem
CDMGB	cable driver modem group buffer
CDN	compressed dial number
CDO	command duty officer
CDOC	counterdrug operations center
CDOPS	counterdrug operations
CDP	commander's dissemination policy; landing craft air cushion departure point
CDR	commander; continuous data recording
CDRAFSOF	commander, Air Force special operations forces
CDRAR	Commander, US Army North
CDRCF	Commander, Combined Forces Command
CDRESC	commander, electronic security command
CDREUDAC	Commander, European Command Defense Analysis Center (ELINT) or European Data Analysis Center
CDRFORSCOM	Commander, Forces Command
CDRG	catastrophic disaster response group (FEMA)
CDRJSOTF	commander, joint special operations task force
CDRL	contract data requirements list
CDRMTMC	Commander, Military Traffic Management Command
CDRNORAD	Commander, North American Aerospace Defense Command
CD-ROM	compact disc read-only memory
CDRTSOC	commander, theater special operations command
CDRUNC	Commander, United Nations Command
CDRUSAINSCOM	Commander, United States Army Intelligence and Security Command
CDRUSCENTCOM	Commander, United States Central Command
CDRUSELEMNORAD	Commander, United States Element, North American Aerospace Defense Command
CDRUSEUCOM	Commander, United States European Command
CDRUSJFCOM	Commander, United States Joint Forces Command
CDRUSNAVEUR	Commander, United States Naval Forces, Europe
CDRUSNORTHCOM	Commander, United States Northern Command
CDRUSPACOM	Commander, United States Pacific Command
CDRUSSOCOM	Commander, United States Special Operations Command
CDRUSSOUTHCOM	Commander, United States Southern Command
CDRUSSTRATCOM	Commander, United States Strategic Command
CDRUSTRANSCOM	Commander, United States Transportation Command

CDS	Chief of Defence Staff (Canada); container delivery system
CDSSC	continuity of operations plan designated successor service chief
CDU	counterdrug update
C-E	communications-electronics
CE	casualty estimation; command element (MAGTF); communications-electronics; core element; counterespionage
CEA	captured enemy ammunition
CEB	combat engineer battalion
CEC	civil engineer corps
CECOM	communications-electronics command
CEDI	commercial electronic data interface
CEDREP	communications-electronics deployment report
CEE	captured enemy equipment
CEF	civil engineering file; common equipment facility
CEG	common equipment group
CEI	critical employment indicator
CEM	combined effects munition
CEMC	communications-electronics management center
CENTRIXS	Combined Enterprise Regional Information Exchange System
CEOI	communications-electronics operating instructions
CEP	cable entrance panel; circular error probable
CEPOD	communications-electronics post-deployment report
CERF	Central Emergency Revolving Fund (UN)
CERFP	CBRNE enhanced response force package
CERP	Commanders' Emergency Response Program
CERT	computer emergency response team
CERTSUB	certain submarine
CES	coast earth station
CESE	civil engineering support equipment; Communications equipment support element
CESG	communications equipment support group
CESO	civil engineer support office
CESPG	civil engineering support plan group; civil engineering support planning generator
CEXC	combined explosives exploitation cell
CF	Canadian forces; causeway ferry; drift error confidence factor
CFA	Committee on Food Aid Policies and Programmes (UN)
CFACC	combined force air component commander
CFB	Canadian forces base
CFC	Combined Forces Command, Korea
CF-COP	counterfire common operational picture
CFL	Contingency Planning Facilities List; coordinated fire line
CFM	cubic feet per minute
CFO	chief financial officer
CFR	Code of Federal Regulations
CFS	CI force protection source
CFSO	counterintelligence force protection source operations
CFST	coalition forces support team
CG	Chairman's guidance; Coast Guard; commanding general; Comptroller General

CGAS	Coast Guard Air Station
CGAUX	Coast Guard Auxiliary
CGC	Coast Guard Cutter
CGCAP	Coast Guard capabilities plan
CGDEFOR	Coast Guard defense force
CGFMFLANT	Commanding General, Fleet Marine Forces, Atlantic
CGFMFPAC	Commanding General, Fleet Marine Forces, Pacific
CGIS	US Coast Guard Investigative Service
CGLSMP	Coast Guard logistic support and mobilization plan
CGRS	common geographic reference system
CGS	common ground station; continental United States ground station
CGUSAREUR	Commanding General, United States Army, Europe
CH	channel; contingency hospital
CH-53	Sea Stallion
CHAMPUS	Civilian Health and Medical Program for the Uniformed Services
CHARC	counterintelligence and human intelligence analysis and requirements cell
CHB	cargo handling battalion
CHCS	composite health care system
CHCSS	Chief, Central Security Service
CHE	cargo-handling equipment; container-handling equipment
CHET	customs high endurance tracker
CHOP	change of operational control
CHPPM	US Army Center for Health Promotion and Preventive Medicine
CHRIS	chemical hazard response information system
CHSTR	characteristics of transportation resources
CHSTREP	characteristics of transportation resources report
CI	civilian internee; counterintelligence
CIA	Central Intelligence Agency
CIAP	Central Intelligence Agency program; central intelligence architecture plan; integrated architecture program; command intelligence architecture plan; command intelligence architecture program
CIAS	counterintelligence analysis section
CIAT	counterintelligence analytic team
CIB	combined information bureau; controlled image base
CIC	combat information center; combat intelligence center (Marine Corps); combined intelligence center; communications interface controller; content indicator code; counterintelligence center
CICA	counterintelligence coordination authority
CICAD	Inter-American Drug Abuse Control Commission
CID	combat identification; combat intelligence division; criminal investigation division
CIDB	common intelligence database
CIDC	Criminal Investigation Division Command
CIE	collaborative information environment
CIEG/CIEL	common information exchange glossary and language
CIFA	counterintelligence field activity
CIG	communications interface group
CIHO	counterintelligence/human intelligence officer
CIIR	counterintelligence information report

CI/KR	critical infrastructure/key resources
CIL	command information library; critical item list
CILO	counterintelligence liaison officer
CIM	compartmented information management
CIMIC	civil-military cooperation
CIN	cargo increment number
CIO	chief information officer; command intelligence officer
CIOTA	counterintelligence operational tasking authority
CIP	communications interface processor; critical infrastructure protection
CIPSU	communications interface processor pseudo line
CIR	continuing intelligence requirement
CIRM	International Radio-Medical Center
CIRV	common interswitch rekeying variable
CIRVIS	communications instructions for reporting vital intelligence sightings
CIS	common item support; Commonwealth of Independent States; communications interface shelter
CISD	critical incident stress debriefing
CISO	counterintelligence staff office; counterintelligence support officer
CITP	counter-IED targeting program
CIV	civilian
CIVPOL	civilian police
CIWG	communications interoperability working group
CJ-4	combined-joint logistics officer
CJATF	commander, joint amphibious task force
CJB	Congressional Justification Book
CJCS	Chairman of the Joint Chiefs of Staff
CJCSAN	Chairman of the Joint Chiefs of Staff Alerting Network
CJCSI	Chairman of the Joint Chiefs of Staff instruction
CJCSM	Chairman of the Joint Chiefs of Staff manual
CJDA	critical joint duty assignment
CJMAB	Central Joint Mortuary Affairs Board
CJMAO	Central Joint Mortuary Affairs Office; Chief, joint mortuary affairs office
CJSOTF	combined joint special operations task force
CJTF	combined joint task force (NATO); commander, joint task force
CJTF-CS	Commander, Joint Task Force - Civil Support
CJTF-NCR	Commander, Joint Task Force - National Capital Region
C-JWICS	Containerized Joint Worldwide Intelligence Communications System
CKT	circuit
CL	class
CLA	landing craft air cushion launch area
CLD	compact laser designator
CLEA	civilian law enforcement agency
C-level	category level
CLF	cantilever lifting frame; combat logistics force; commander, landing force
CLG	combat logistics group
CLGP	cannon-launched guided projectile
CLIPS	communications link interface planning system
CLPSB	Combatant Commander Logistic Procurement Support Board

CLS	contracted logistic support
CLSS	combat logistic support squadron
CLT	combat lasing team
CLZ	craft landing zone; cushion landing zone
CM	Chairman's memorandum; collection manager; configuration management; consequence management; control modem; countermine
Cm	mean coverage factor
cm	centimeter
CMA	collection management authority
CMAA	Cooperative Military Airlift Agreement
CMAH	commander of a combatant command's Mobile Alternate Headquarters
CMAT	consequence management advisory team
CMC	Commandant of the Marine Corps; crew management cell
Cmc	midpoint compromise coverage factor
CMD	command; cruise missile defense
CMHT	consequence management home team
CMMA	collection management mission application
CMO	Central Measurement and Signature Intelligence (MASINT) Organization; chief medical officer; chief military observer; civil-military operations; collection management office(r); configuration management office
CMOC	Cheyenne Mountain Operations Center; civil-military operations center
CMOS	cargo movement operations system; Complementary metal-oxide semiconductor
CMP	communications message processor
CMPF	commander, maritime pre-positioned force
CMPT	consequence management planning team
CM R&A	consequence management response and assessment
CMRT	consequence management response team
CMS	cockpit management system; command management system; community management staff; community security materiel system; contingency mutual support; crisis management system
CMST	consequence management support team
CMTS	comments
CMTU	cartridge magnetic tape unit
CMV	commercial motor vehicle
CMX	crisis management exercise
CN	counternarcotic
CNA	computer network attack
CNAC	Customs National Aviation Center (USCS)
CNASP	chairman's net assessment for strategic planning
CNC	Counter-Narcotics Center (CIA); Crime and Narcotics Center
CNCE	communications nodal control element
CND	computer network defense; counternarcotics division
CNE	computer network exploitation; Counter Narcotics Enforcement
CNGB	Chief, National Guard Bureau
CNO	Chief of Naval Operations; computer network operations
CNOG	Chairman, Nuclear Operations Group
CNRF	Commander, Naval Reserve Forces

CNSG	Commander, Naval Security Group
CNTY	country
CNWDI	critical nuclear weapons design information
CO	commanding officer
COA	course of action
COAA	course-of-action analysis
COAMPS	Coupled Ocean Atmosphere Mesoscale Prediction System
COB	collocated operating base; contingency operating base
COBOL	common business-oriented language
COC	combat operations center
CoC	Code of Conduct
COCOM	combatant command (command authority)
COD	carrier onboard delivery; combat operations division
COE	Army Corps of Engineers; common operating environment; concept of employment
COF	conduct of fire
COFC	container on flatcar
COG	center of gravity; continuity of government
COGARD	Coast Guard
COI	contact of interest
COIN	counterinsurgency
COLDS	cargo offload and discharge system
COLISEUM	community on-line intelligence system for end-users and managers
COLT	combat observation and lasing team
COM	chief of mission; collection operations management; command; commander
COMACC	Commander, Air Combat Command
COMAFFOR	commander, Air Force forces
COMAFSOC	Commander, Air Force Special Operations Command
COMAJF	commander, allied joint force
COMALF	commander airlift forces
COMALOC	commercial air line of communications
COMARFOR	commander, Army forces
COMCAM	combat camera
COMCARGRU	commander, carrier group
COMCEN	communications center
COMCRUDESGRU	Commander, cruiser destroyer group
COMDCAEUR	Commander, Defense Communications Agency Europe
COMDESRON	Commander destroyer squadron
COMDT COGARD	Commandant, United States Coast Guard
COMDTINST	Commandant, United States Coast Guard instruction
COMICEDEFOR	Commander, United States Forces, Iceland
COMIDEASTFOR	Commander, Middle East Forces
COMINEWARCOM	Commander, Mine Warfare Command
COMINT	communications intelligence
COMJCSE	Commander, Joint Communications Support Element
COMJIC	Commander, Joint Intelligence Center
COMJSOTF	commander, joint special operations task force
COMLANDFOR	Commander, land forces
COMLANTAREACOGARD	Commander, Coast Guard Atlantic Area

COMLOGGRU	combat logistics group
COMM	communications
COMMARFOR	commander, Marine Corps forces
COMMARFORNORTH	Commander, Marine Corps Forces North
COMMDZ	Commander, Maritime Defense Zone
COMMZ	communications zone
COMNAV	Committee for European Airspace Coordination Working Group on Communications and Navigation Aids
COMNAVAIRLANT	Commander, Naval Air Force, Atlantic
COMNAVAIRPAC	Commander, Naval Air Force, Pacific
COMNAVAIRSYSCOM	Commander, Naval Air Systems Command
COMNAVCOMTELCOM	Commander, Naval Computer and Telecommunications Command
COMNAVFOR	Commander, Navy forces
COMNAVMETOCCOM	Commander, Naval Meteorology and Oceanography Command
COMNAVSEASYSCOM	Commander, Naval Sea Systems Command
COMNAVSECGRP	Commander, United States Navy Security Group
COMNAVSURFLANT	Commander, Naval Surface Force, Atlantic
COMNAVSURFPAC	Commander, Naval Surface Force, Pacific
COMP	component
COMPACAF	Commander, Pacific Air Forces
COMPACAREACOGARD	Commander, Coast Guard Pacific Area
COMPACFLT	Commander, Pacific Fleet
COMPASS	common operational modeling, planning, and simulation strategy; Computerized Movement Planning and Status System
COMPES	contingency operations mobility planning and execution system
COMPLAN	communications plan
COMPUSEC	computer security
COMSAT	communications satellite
COMSC	Commander, Military Sealift Command
COMSCINST	Commander, Military Sealift Command instruction
COMSEC	communications security
COMSOC	Commander, Special Operations Command
COMSOCCENT	Commander, Special Operations Command, United States Central Command
COMSOCEUR	Commander, Special Operations Command, United States European Command
COMSOCPAC	Commander Special Operations Command, United States Pacific Command
COMSOCSOUTH	Commander Special Operations Command, United States Southern Command
COMSOF	commander, special operations forces
COMSTAT	communications status
COMSUBLANT	Commander Submarine Force, United States Atlantic Fleet
COMSUBPAC	Commander Submarine Force, United States Pacific Fleet
COMSUPNAVFOR	commander, supporting naval forces
COMTAC	tactical communications
COMUSAFE	Commander, United States Air Force in Europe
COMUSARCENT	Commander, United States Army Forces, Central Command
COMUSCENTAF	Commander, United States Air Force, Central Command
COMUSFLTFORCOM	Commander, US Fleet Forces Command

COMUSFORAZ	Commander, United States Forces, Azores
COMUSJ	Commander, United States Forces, Japan
COMUSK	Commander, United States Forces, Korea
COMUSLANTFLT	Commander, US Atlantic Fleet
COMUSMARCENT	Commander, United States Marine Forces, Central Command
COMUSNAVCENT	Commander, United States Navy, Central Command
COMUSPACFLT	Commander, US Pacific Fleet
COMUSSOCJFCOM	Commander Special Operations Command, United States Joint Forces Command
CONCAP	construction capabilities contract (Navy); Construction Capabilities Contract Process; construction capabilities contract program
CONEX	container express
CONEXPLAN	contingency and exercise plan
CONOPS	concept of operations
CONPLAN	concept plan; operation plan in concept format
CONR	continental United States North American Aerospace Defense Command Region
CONTRAIL	condensation trail
CONUS	continental United States
CONUSA	Continental United States Army
COOP	continuity of operations
COP	common operational picture
COP-CSE	common operational picture-combat support enabled
COPE	custodian of postal effects
COPG	chairman, operations planners group
COPPERHEAD	name for cannon-launched guided projectile
COPS	communications operational planning system
COR	contracting officer representative
CORE	contingency response program
COS	chief of staff; chief of station; critical occupational specialty
COSCOM	corps support command
COSMIC	North Atlantic Treaty Organization (NATO) security category
COSPAS	*cosmicheskaya sistyema poiska avariynch sudov*–Space system for search of distressed vessels (Russian satellite system)
COSR	combat and operational stress reactions
COT	commanding officer of troops; crisis operations team
COTHEN	Customs Over-the Horizon Enforcement Network
COTP	captain of the port
COTS	cargo offload and transfer system; commercial off-the-shelf; container offloading and transfer system
COU	cable orderwire unit
counter C3	counter command, control, and communications
COVCOM	covert communications
CP	check point; collection point; command post; contact point; control point; counterproliferation
CP&I	coastal patrol and interdiction
CPA	Chairman’s program assessment; closest point of approach
CPD	combat plans division
CPE	customer premise equipment
CPFL	contingency planning facilities list

CPG	central processor group; Commander, Amphibious Group; Contingency Planning Guidance
CPI	crash position indicator
CPIC	coalition press information center
CPM	civilian personnel manual
CPO	chief petty officer; complete provisions only
CPR	cardiopulmonary resuscitation; Chairman's program recommendation
CPRC	coalition personnel recovery center
CPS	characters per second; collective protective shelter
CPT	common procedural terminology
CPU	central processing unit
CPX	command post exercise
CRA	command relationships agreement; continuing resolution authority; coordinating review authority
CRAF	Civil Reserve Air Fleet
CRAM	control random access memory
CRB	configuration review board
CRC	circuit routing chart; control and reporting center; CONUS replacement center; COOP response cell
CRD	cyclic redundancy rate capstone requirements document; Chemical reconnaissance detachment; combatant commander's required date
CRE	contingency response element; control reporting element
CREST	casualty and resource estimation support tool
CRF	channel reassignment function
CRG	contingency response group
CRI	collective routing indicator
CRIF	cargo routing information file
CRITIC	critical information; critical intelligence communication; critical message (intelligence)
CRITICOMM	critical intelligence communications system
CRM	collection requirements management; crew resource management
CrM	crisis management
CRO	combat rescue officer
CROP	common relevant operational picture
CRP	control and reporting post
CRRC	combat rubber raiding craft
CRS	Catholic Relief Services; Chairman's readiness system; coastal radio station; community relations service; container recovery system
CRSP	centralized receiving and shipping point
CRT	cathode ray tube
CRTS	casualty receiving and treatment ship
CR-UAV	close-range unmanned aerial vehicle
CRW	contingency response wing
CRYPTO	cryptographic
CS	call sign; Chaplain Service (Air Force); circuit switch; civil support; coastal station; combat support; content staging; controlled space; creeping line single-unit; critical source
CSA	Chief of Staff, United States Army; combat support agency; container stuffing activity
CSAAS	combat support agency assessment system
CSADR	combat support agency director's report

CSAF	Chief of Staff, United States Air Force
CSAM	computer security for acquisition managers
CSAR	combat search and rescue
CSAR3	combat support agency responsiveness and readiness report
CSARTE	combat search and rescue task element
CSARTF	combat search and rescue task force
CSB (ME)	combat support brigade (maneuver enhancement)
CSC	combat support center; community support center; convoy support center; creeping line single-unit coordinated; International Convention for Safe Containers
CSCC	coastal sea control commander
CSE	client server environment; combat support enhanced; combat support equipment
CSEL	circuit switch select line; combat survivor evader locator; command senior enlisted leader
CSEP	Chairman of the Joint Chiefs of Staff -sponsored exercise program
CSG	carrier strike group; Chairman's Staff Group; Coordinating subgroup; cryptologic services group; Cryptologic Support Group
CSGN	coordinating subgroup for narcotics
CSH	combat support hospital
CSI	critical safety item; critical sustainability item
CSIF	communications service industrial fund
CSIPG	circuit switch interface planning guide
CSL	combat stores list; cooperative security location
CSNP	causeway section, nonpowered
CSNP(BE)	causeway section, nonpowered (beach end)
CSNP(I)	causeway section, nonpowered (intermediate)
CSNP(SE)	causeway section, nonpowered (sea end)
CSO	communications support organization
CSOA	combined special operations area
CSOB	command systems operations branch
CSOD	command systems operation division
CSP	call service position; career sea pay; causeway section, powered; commence search point; contracting support plan; crisis staffing procedures (JCS); cryptologic support package
CSPAR	combatant commander's preparedness assessment report
CSR	central source registry; combatant commander's summary report; commander's summary report; controlled supply rate
CSRF	common source route file
CSS	central security service; combat service support; communications subsystem; coordinator surface search
CSSA	combat service support area
CSSC	coded switch set controller
CSSE	combat service support element (MAGTF)
CSST	combat service support team
CSSU	combat service support unit
CST	customer service team
CSW	compartment stowage worksheet; coordinate seeking weapons
CT	communications terminal; control telemetry; counterterrorism; country team
CTA	common table of allowance

CTL	candidate target list
CTAPS	contingency Theater Air Control System automated planning system
CTC	cargo transfer company (USA); counterterrorist center
CTF	combined task force
CTG	commander, task group
CTID	communications transmission identifier
CTM	core target material
CTOC	corps tactical operations center
CTP	common tactical picture
CTRIF	Combating Terrorism Readiness Initiative Fund
CTS	Commodity Tracking System
CTSS	central targeting support staff
CTU	commander, task unit
CU	cubic capacity; common unit
CUL	common-user logistics
CULT	common-user land transportation
CV	aircraft carrier; carrier; curriculum vitae
CVAMP	Core Vulnerability Assessment Management Program
CVBG	carrier battle group
CVISC	combat visual information support center
CVN	aircraft carrier, nuclear
CVR	cockpit voice recorder
CVS	commercial vendor services
CVSD	continuous variable slope delta
CVT	criticality-vulnerability-threat
CVW	carrier air wing; cryptovariable weekly (GPS)
CVWC	carrier strike group air wing commander
CW	carrier wave; chemical warfare; continuous wave
CWC	Chemical Weapons Convention; composite warfare commander
CWDE	chemical warfare defense equipment
CWO	communications watch officer
CWP	causeway pier
CWPD	Conventional War Plans Division, Joint Staff (J-7)
CWR	calm water ramp
CWT	combat weather team; customer wait time
CY	calendar year
D	total drift, data
d	surface drift
D&D	denial and deception
D&F	determinations and findings
D&M	detection and monitoring
D&R	debrief and reintegrate
D3A	decide, detect, deliver, and assess
D/A	digital-to-analog
DA	data adapter aerospace drift; data administrator; Department of the Army; Development Assistance; direct action; Directorate for Administration (DIA); double agent
Da	aerospace drift
DA&M	Director of Administration and Management

DAA	designated approving authority; display alternate area routing lists
DAADC(AMD)	deputy area air defense commander for air and missile defense
DAAS	defense automatic addressing system
DAASO	defense automatic addressing system office
DAB	Defense Acquisition Board
DAC	Defense Intelligence Agency (DIA) counterintelligence and security activity; Department of Army civilians
DACB	data adapter control block
DACG	departure airfield control group
DACM	data adapter control mode
DADCAP	dawn and dusk combat air patrol
DAF	Department of the Air Force
DAFL	directive authority for logistics
DAICC	domestic air interdiction coordinator center
DAL	defended asset list
DALIS	Disaster Assistance Logistics Information System
DALS	downed aviator locator system
DAMA	demand assigned multiple access
DAMES	defense automatic addressing system (DAAS) automated message exchange system
DAN	Diver's Alert Network
DAO	defense attaché office; defense attaché officer; department/agency/organization
DAP	designated acquisition program
DAR	distortion adaptive receiver
DARO	Defense Airborne Reconnaissance Office
DARPA	Defense Advanced Research Projects Agency
DART	disaster assistance response team; downed aircraft recovery team; dynamic analysis and replanning tool
DAS	deep air support (USMC); defense attaché system; direct access subscriber; direct air support
DAS3	decentralized automated service support system
DASA	Department of the Army (DA) staff agencies
DASC	direct air support center
DASC(A)	direct air support center (airborne)
DASD	Deputy Assistant Secretary of Defense
DASD-CN	Deputy Assistant Secretary of Defense for Counternarcotics
DASD(H&RA)	Deputy Assistant Secretary of Defense (Humanitarian & Refugee Affairs)
DASD(I)	Deputy Assistant Secretary of Defense (Intelligence)
DASD(PK/HA)	Deputy Assistant Secretary of Defense (Peacekeeping and Humanitarian Affairs)
DASD(S&IO)	Deputy Assistant Secretary of Defense (Security and Information Operations)
DASSS	decentralized automated service support system
DAT	deployment action team
DATT	defense attaché
DATU	data adapter termination unit
dB	decibel
DBA	database administrator

DBDB	digital bathymetric database
DBG	database generation
DBI	defense budget issue
DBMS	database management system; Defense-Business Management System
DBOF	Defense Business Operations Fund
DBSS	Defense Blood Standard System
DBT	design basis threat
D/C	downconverter
DC	Deputies Committee; direct current; dislocated civilian
DCA	Defense Communications Agency; Defense Cooperation Agreements; defensive counterair; dual-capable aircraft
DCAA	Defense Contract Audit Agency
DCAM	Defense Medical Logistics Standard Support (DMLSS) customer assistance module
DCC	damage control center; deployment control center
DCCC	defense collection coordination center
DCCEP	developing country combined exercise program
DCD	data collection device
DCE	defense coordinating element
D-cell	deployment cell
DCGS	Distributed Common Ground/Surface System
DCI	defense critical infrastructure; Director of Central Intelligence; dual channel interchange
D/CI&SP	Director, Counterintelligence and Security Programs
DCID	Director of Central Intelligence directive
DCIIS	Defense Counterintelligence Information System
DCIO	defense criminal investigative organization
DCIP	Defense Critical Infrastructure Program
DCIS	Defense Criminal Investigative Services
DCJTF	deputy commander, joint task force
DCM	data channel multiplexer; deputy chief of mission
DCMA	Defense Contract Management Agency
DCMC	Office of Deputy Chairman, Military Committee
DCMO	deputy chief military observer
DCN	data link coordination net
DCNO	Deputy Chief of Naval Operations
DCO	defense coordinating officer (DOD); dial central office
DCP	Defense Contingency Program
DCPA	Defense Civil Preparedness Agency
DCPG	digital clock pulse generator
DCR	DOTMLPF change recommendation
DCS	Defense Communications System; Defense Courier Service; deputy chief of staff; digital computer system
DCSCU	dual capability servo control unit
DC/S for RA	Deputy Chief of Staff for Reserve Affairs
DCSINT	Deputy Chief of Staff for Intelligence
DCSLOG	Deputy Chief of Staff for Logistics, US Army
DCSOPS	Deputy Chief of Staff for Operations and Plans, United States Army
DCSPER	Deputy Chief of Staff for Personnel, United States Army

DCST	Defense Logistics Agency (DLA) contingency support team
DCTS	Defense Collaboration Tool Suite
DD	Department of Defense (form); destroyer (Navy ship)
DDA	Deputy Director for Administration (CIA); esignated development activity
D-day	unnamed day on which operations commence or are scheduled to commence
DDC	data distribution center; defense distribution center
DDCI	Deputy Director of Central Intelligence (CIA)
DDCI/CM	Deputy Director of Central Intelligence for Community Management
DDG	guided missile destroyer
DDI	Deputy Director of Intelligence (CIA)
DDL	digital data link
DDM	digital data modem
DDMA	Defense Distribution Mapping Activity
DDMS	Deputy Director for Military Support
DDO	Deputy Director of Operations (CIA)
DDOC	deployment and distribution operations center
DDP	detailed deployment plan
DDR&E	director of defense research and engineering
DDRRR	disarmament, demobilization, repatriation, reintegration, and resettlement
DDS	defense dissemination system; Deployable Disbursing System; dry deck shelter
DDSM	Defense Distinguished Service Medal
DDS&T	Deputy Director for Science & Technology (CIA)
DDWSO	Deputy Director for Wargaming, Simulation, and Operations
DE	damage expectancy; delay equalizer; directed energy
De	total drift error
de	individual drift error
DEA	Drug Enforcement Administration
dea	aerospace drift error
DEACN	Drug Enforcement Administration Communications Network
DEAR	disease and environmental alert report
DEARAS	Department of Defense (DOD) Emergency Authorities Retrieval and Analysis System
DeCA	Defense Commissary Agency
DECL	declassify
DEFCON	defense readiness condition
DEFSMAC	Defense Special Missile and Astronautics Center
DEL	deployable equipment list
DEMARC	demarcation
de max	maximum drift error
DEMIL	demilitarization
de min	minimum drift error
de minimax	minimax drift error
DeMS	deployment management system
DEMUX	demultiplex
DEP	Delayed Entry Program; deployed
DEP&S	Drug Enforcement Plans and Support

DEPCJTF	deputy commander, joint task force
DEPID	deployment indicator code
DEPMEDS	deployable medical systems
DepOpsDeps	Service deputy operations deputies
DEPORD	deployment order
DESC	Defense Energy Support Center
DESCOM	Depot System Command (Army)
DESIGAREA	designated area message
DEST	destination; domestic emergency support team
DET	detachment; detainee
DETRESFA	distress phase (ICAO)
DEW	directed-energy warfare
DF	direction finding; dispersion factor; disposition form
DFARS	Defense Federal Acquisition Regulation Supplement
DFAS	Defense Finance and Accounting Service
DFAS-DE	Defense Finance and Accounting Service-Denver
DFC	deputy force commander
DFE	Defense Joint Intelligence Operations Center forward element; division force equivalent
DFM	deterrent force module
DFO	disaster field office (FEMA)
DFR	Defense Fuel Region
DFR/E	Defense Fuel Region, Europe
DFRIF	Defense Freight Railway Interchange Fleet
DFR/ME	Defense Fuel Region, Middle East
DFSC	Defense Fuel Supply Center
DFSP	Defense Fuel Support Point
DFT	deployment for training
DG	defense guidance
DGIAP	Defense General Intelligence and Applications Program
DGM	digital group multiplex
DGZ	desired ground zero
DH	death due to hostilities; Directorate for Human Intelligence (DIA)
DHE	Department of Defense (DOD) human intelligence (HUMINT) element
DHHS	Department of Health and Human Services
DHM	Department of Defense human intelligence manager
DHMO	Department of Defense human intelligence management office
DHS	Defense Human Intelligence (HUMINT) Service; Department of Homeland Security; Director of Health Services
DI	Defense Intelligence Agency (DIA) Directorate for Analysis
DIA	Directorate for Intelligence Production; Discrete identifier; dynamic interface
DIA	Defense Intelligence Agency
DIAC	Defense Intelligence Analysis Center
DIA/DHX	Defense Intelligence Agency, Directorate of Human Intelligence, Office of Document and Media Operations
DIAM	Defense Intelligence Agency manual; Defense Intelligence Agency memorandum

DIAP	Defense Intelligence Analysis Program; Drug Interdiction Assistance Program
DIAR	Defense Intelligence Agency (DIA) regulation
DIB	defense industrial base
DIBITS	digital in-band interswitch trunk signaling
DIBRS	defense incident-based reporting system
DIBTS	digital in-band trunk signaling
DICO	Data Information Coordination Office
DIDHS	Deployable Intelligence Data Handling System
DIDO	designated intelligence disclosure official
DIDS	Defense Intelligence Dissemination System
DIEB	Defense Intelligence Executive Board
DIEPS	Digital Imagery Exploitation Production System
DIG	digital
DIGO	Defence Imagery and Geospatial Organisation
DII	defense information infrastructure
DII-COE	defense information infrastructure-common operating environment
DIILS	Defense Institute of International Legal Studies
DIJE	Defense Intelligence Joint Environment
DILPA	diphase loop modem-A
DIMA	drilling individual mobilization augmentee
DIN	defense intelligence notice
DINET	Defense Industrial Net
DINFOS	Defense Information School
DIOC	drug interdiction operations center
DIPC	defense industrial plant equipment center
DIPFAC	diplomatic facility
DIPGM	diphase supergroup modem
DIRINT	Director of Intelligence (USMC)
DIRJIATF	director, joint inter-agency task force
DIRLAUTH	direct liaison authorized
DIRM	Directorate for Information and Resource Management
DIRMOBFOR	director of mobility forces
DIRNSA	Director of National Security Agency
DIS	daily intelligence summary; defense information system; Defense Investigative Service; distributed interactive simulation
DISA	Defense Information Systems Agency
DISA-LO	Defense Information Systems Agency - liaison officer
DISANMOC	Defense Information Systems Agency Network Management and Operations Center
DISCOM	division support command (Army)
DISGM	diphase supergroup
DISN	Defense Information Systems Network
DISN-E	Defense Information Systems Network – Europe
DISO	defense intelligence support office
DISP	drug investigation support program (FAA)
DISUM	daily intelligence summary
DITDS	defense information threat data system; defense intelligence threat data system

DITSUM	defense intelligence terrorist summary
DJIOC	Defense Joint Intelligence Operations Center
DJS	Director, Joint Staff
DJSM	Director, Joint Staff memorandum
DJTFAC	deployable joint task force augmentation cell
DJTFS	deputy joint task force surgeon
DLA	Defense Logistics Agency
DLAM	Defense Logistics Agency manual
DLAR	Defense Logistics Agency regulation
DLEA	drug law enforcement agency
DLED	dedicated loop encryption device
DLIS	Defense Logistics Information Service
DLP	data link processor
DLPMA	diphase loop modem A
DLQ	deck landing qualification
DLR	depot-level repairable
DLSA	Defense Legal Services Agency
DLSS	Defense Logistics Standard Systems
DLTM	digital line termination module
DLTU	digital line termination unit
DM	detection and monitoring
dmax	maximum drift distance
DMB	datum marker buoy
DMC	data mode control
DMD	digital message device
DMDC	defense management data center; defense manpower data center
DME	distance measuring equipment
DMI	director military intelligence
DMIGS	Domestic Mobile Integrated Geospatial-Intelligence System
dmin	minimum drift distance
DML	data manipulation language
DMLSS	Defense Medical Logistics Standard Support
DMO	directory maintenance official
DMOS	duty military occupational specialty
DMPI	designated mean point of impact; desired mean point of impact
DMRD	defense management resource decision
DMRIS	defense medical regulating information system
DMS	defense message system; defense meteorological system; director of military support
DMSB	Defense Medical Standardization Board
DMSM	Defense Meritorious Service Medal
DMSO	Defense Modeling and Simulation Office; director of major staff office; Division Medical Supply Office
DMSP	Defense Meteorological Satellite Program
DMSSC	defense medical systems support center
DMT	disaster management team (UN)
DMU	disk memory unit
DMZ	demilitarized zone
DN	digital nonsecure
DNA	Defense Nuclear Agency; deoxyribonucleic acid

DNAT	defense nuclear advisory team
DNBI	disease and nonbattle injury
DNBI casualty	disease and nonbattle injury casualty
DNC	digital nautical chart
DND	Department of National Defence
DNDO	Domestic Nuclear Detection Office
DNGA	Director of National Geospatial-Intelligence Agency
DNI	Director of National Intelligence; Director of Naval Intelligence
DNIF	duty not involving flying
DNMSP	driftnet monitoring support program
DNSO	Defense Network Systems Organization
DNVT	digital nonsecure voice terminal
DNY	display area code (NYX) routing
DOA	dead on arrival; director of administration
DOB	date of birth; dispersal operating base
DOC	Department of Commerce; designed operational capability
DOCC	deep operations coordination cell
DOCDIV	documents division
DOCEX	document exploitation
DOCNET	Doctrine Networked Education and Training
DOD	Department of Defense
DODAAC	Department of Defense activity address code
DODAAD	Department of Defense Activity Address Directory
DODAC	DOD ammunition code
DODD	Department of Defense directive
DODDS	Department of Defense Dependent Schools
DODEX	Department of Defense intelligence system information system extension
DODFMR	Department of Defense Financial Management Regulation
DODI	Department of Defense instruction
DODIC	Department of Defense identification code
DODID	Department of Defense Intelligence Digest
DODIIS	Department of Defense Intelligence Information System
DODIPC	Department of Defense intelligence production community
DODIPP	Department of Defense Intelligence Production Program
DOD-JIC	Department of Defense Joint Intelligence Center
DODM	data orderwire diphase modem
DOE	Department of Energy
DOF	degree of freedom
DOI	Defense Special Security Communications System (DSSCS) Operating Instructions; Department of Interior
DOJ	Department of Justice
DOL	Department of Labor
DOM	day of month
DOMS	director of military support
DON	Department of the Navy
DOPMA	Defense Officer Personnel Management Act
DOR	date of rank
DOS	date of separation; days of supply; denial of service; Department of State; disk operating system

DOT	Department of Transportation
DOTEO	Department of Transportation emergency organization
DOTMLPF	doctrine, organization, training, materiel, leadership and education, personnel, and facilities
DOW	data orderwire; died of wounds
DOX-T	direct operational exchange-tactical
DOY	day of year
DP	Air Force component plans officer (staff); decisive point; Directorate for Policy Support (DIA); displaced person
dp	parachute drift
DPA	Defense Production Act
DPAS	Defense Priorities and Allocation System
DPC	deception planning cell; Defense Planning Committee (NATO)
DPEC	displaced person exploitation cell
DPG	Defense Planning Guidance
DPI	desired point of impact
dpi	dots per inch
DPICM	dual purpose improved conventional munitions
DPKO	Department of Peacekeeping Operations
DPLSM	dipulse group modem
DPM	dissemination program manager
DPMO	Defense Prisoner of War/Missing Personnel Office
DPO	distribution process owner
DPP	data patch panel; distributed production program
DPPDB	digital point positioning database
DPQ	defense planning questionnaire (NATO)
DPR	display non-nodal routing
DPRB	Defense Planning and Resources Board
DPRE	displaced persons, refugees, and evacuees
DPS	data processing system
DPSC	Defense Personnel Support Center
DPSK	differential phase shift keying
DR	dead reckoning; digital receiver; disaster relief
DRB	Defense Resources Board
DRe	dead reckoning error
DRMD	deployments requirements manning document
DRMO	Defense Reutilization and Marketing Office
DRMS	Defense Reutilization and Marketing Service; Distance root-mean-square
DRO	departmental requirements office
DRSN	Defense Red Switched Network
DRT	dead reckoning tracer
DRTC	designated reporting technical control
DS	Directorate for Information Systems and Services (DIA); direct support; doctrine sponsor
DSA	defense special assessment (DIA); defensive sea area
DSAA	Defense Security Assistance Agency
DSAR	Defense Supply Agency regulation
DSB	digital in-band trunk signaling (DIBTS) signaling buffer
DSC	digital selective calling

DSCA	Defense Security Cooperation Agency; defense support of civil authorities
DSCP	Defense Supply Center Philadelphia
DSCR	Defense Supply Center Richmond
DSCS	Defense Satellite Communications System
DSCSOC	Defense Satellite Communications System operations center
DSDI	digital simple data interface
DSG	digital signal generator
DSI	defense simulation internet
DSL	display switch locator (SL) routing
DSMAC	digital scene-matching area correlation
DSN	Defense Switched Network
DSNET	Defense Secure Network
DSNET-2	Defense Secure Network-2
DSO	defensive systems officer
DSOE	deployment schedule of events
DSP	Defense Satellite Program; Defense Support Program
DSPD	defense support to public diplomacy
DSPL	display system programming language
DSPS	Director, Security Plans and Service
DSR	defense source registry
DSS	Defense Security Service; Distribution Standard System
DSSCS	Defense Special Security Communications System
DSSM	Defense Superior Service Medal
DSSO	data system support organization; defense sensitive support office; defense systems support organization
DSSR	Department of State Standardized Regulation
DST	deployment support team
DSTP	Director of Strategic Target Planning
DSTR	destroy
DSTS-G	DISN Satellite Transmission Services - Global
DSVL	doppler sonar velocity log
DSVT	digital subscriber voice terminal
DT	Directorate for MASINT and Technical Collection (DIA)
DTA	Defense Threat Assessment; dynamic threat assessment
DTAM	defense terrorism awareness message
DTD	detailed troop decontamination
DTE	data terminal equipment; developmental test and evaluation
DTED	digital terrain elevation data
DTG	date-time group; digital trunk group (digital transmission group)
DTIP	Disruptive Technology Innovations Partnership (DIA)
DTL	designator target line
DTMF	dual tone multi-frequency
DTMR	defense traffic management regulation
DTO	division transportation office; drug trafficking organization
DTOC	division tactical operations center
DTR	defense transportation regulation
DTRA	Defense Threat Reduction Agency
DTRACS	Defense Transportation Reporting and Control System
DTRATCA	Defense Threat Reduction and Treaty Compliance Agency

DTS	Defense Transportation System; Defense Travel System; diplomatic telecommunications service
DTTS	Defense Transportation Tracking System
DUSD (CI&S)	Deputy Under Secretary of Defense for Counterintelligence and Security
DUSDL	Deputy Under Secretary of Defense for Logistics
DUSDP	Deputy Under Secretary of Defense for Policy
DUSTWUN	duty status-whereabouts unknown
DV	distinguished visitor
DVA	Department of Veterans Affairs
DVD	digital video device; digital video disc; direct vendor delivery
DVITS	Digital Video Imagery Transmission System
DVOW	digital voice orderwire
DVT	deployment visualization tool
DWAS	Defense Working Capital Accounting System
DWMCF	double-wide modular causeway ferry
DWRIA	died of wounds received in action
DWT	deadweight tonnage
DWTS	Digital Wideband Transmission System
DX	direct exchange; Directorate for External Relations (DIA)
DZ	drop zone
DZC	drop zone controller
DZCO	drop zone control officer
DZSO	drop zone safety officer
DZST	drop zone support team
DZSTL	drop zone support team leader
E	total probable error
E&DCP	evaluation and data collection plan
E&E	emergency and extraordinary expense authority; Evasion and escape
E&EE	emergency and extraordinary expense
E&I	engineering and installation
E&M	ear and mouth; special signaling leads
E1	Echelon 1
E2	Echelon 2
E3	Echelon 3; electromagnetic environmental effects
E4	Echelon 4
E5	Echelon 5
E-8C	joint surveillance, target attack radar system (JSTARS) aircraft
EA	electronic attack; emergency action; evaluation agent; executive agent; executive assistant
EAC	echelons above corps (Army); emergency action; emergency action committee
EACS	expeditionary aeromedical evacuation crew member support
EACT	expeditionary aeromedical evacuation coordination team
EAD	earliest arrival date; echelons above division (Army); extended active duty
EADS	Eastern Air Defense Sector
EAES	expeditionary aeromedical evacuation squadron
EAF	expeditionary aerospace forces

EAI	executive agent instruction
EALT	earliest anticipated launch time
EAM	emergency action message
EAP	emergency action plan; emergency action procedures
EAP-CJCS	emergency action procedures of the Chairman of the Joint Chiefs of Staff
EARLY	evasion and recovery supplemental data report
E-ARTS	en route automated radar tracking system
EASF	expeditionary aeromedical staging facility
EAST	expeditionary aeromedical evacuation staging team
EASTPAC	eastern Pacific Ocean
EBCDIC	extended binary coded decimal interchange code
EBS	environmental baseline survey
EC	electronic combat; enemy combatant; error control; European Community
ECAC	Electromagnetic Compatibility Analysis Center
ECB	echelons corps and below (Army)
ECC	engineer coordination cell; evacuation control center
ECHA	Executive Committee for Humanitarian Affairs
ECM	electronic countermeasures
ECN	electronic change notice; Minimum Essential Emergency Communications Network
ECO	electronic combat officer
ECOSOC	Economic and Social Council (UN)
ECP	emergency command precedence; engineering change proposal; entry control point
ECS	expeditionary combat support
ECU	environmental control unit
ED	envelope delay; evaluation directive
EDA	Excess Defense Articles
EDC	estimated date of completion
EDD	earliest delivery date
EDI	electronic data interchange
EDSS	equipment deployment and storage system
EE	emergency establishment
EEA	environmental executive agent
EEBD	emergency escape breathing device
EECT	end evening civil twilight
EED	electro-explosive device; emergency-essential designation
EEE	emergency and extraordinary expense
EEFI	essential elements of friendly information
EEI	essential element of information
EELV	evolved expendable launch vehicle
EEO	equal employment opportunity
EEPROM	electronic erasable programmable read-only memory
EER	enlisted employee review; extended echo ranging
EEZ	exclusive economic zone
EFA	engineering field activity
EFAC	emergency family assistance center
EFD	engineering field division

EFST	essential fire support task
EFT	electronic funds transfer
EFTO	encrypt for transmission only
EGM	Earth Gravity Model
EGS	Earth ground station
EH	explosive hazard
EHCC	explosive hazards coordination cell
EHF	extremely high frequency
EHO	environmental health officer
EHRA	environmental health risk assessment
EHSA	environmental health site assessment
EHT	explosive hazard team
EI	environmental information; exercise item
EIA	Electronic Industries Association
EIS	Environmental Impact Statement
ELBA	emergency locator beacon
ELCAS	elevated causeway system
ELCAS(M)	elevated causeway system (modular)
ELCAS(NL)	elevated causeway system (Navy lighterage)
ELD	emitter locating data
ELECTRO-OPTINT	electro-optical intelligence
ELINT	electronic intelligence
ELIST	enhanced logistics intratheater support tool
ELOS	extended line of sight
ELPP	equal level patch panel
ELR	extra-long-range aircraft
ELSEC	electronics security
ELSET	element set
ELT	emergency locator transmitter
ELV	expendable launch vehicle
ELVA	emergency low visibility approach
EM	electromagnetic; executive manager
EMAC	emergency management assistance compact
E-mail	electronic mail
EMALL	electronic mall
EMC	electromagnetic compatibility
EMCON	emissions control EMCON orders
EMD	effective miss distance
EME	electromagnetic environment
EMEDS	Expeditionary Medical Support
EMF	expeditionary medical facility
EMI	electromagnetic interface; electromagnetic interference
EMIO	expanded maritime interception operations
EMP	electromagnetic pulse
EMR hazards	electromagnetic radiation hazards
EMS	electromagnetic spectrum; emergency medical services
EMSEC	emanations security
EMT	emergency medical technician; emergency medical treatment
EMV	electromagnetic vulnerability
ENCOM	engineer command (Army)

ENDEX	exercise termination
ENL	enlisted
ENSCE	enemy situation correlation element
ENWGS	Enhanced Naval Warfare Gaming System
EO	electro-optical; end office; equal opportunity; executive order; eyes only
EOB	electronic order of battle; enemy order of battle
EOC	early operational capability; emergency operations center
EOD	explosive ordnance disposal
EOI	electro-optic(al) imagery
EO-IR	electro-optical-infrared
EO-IR CM	electro-optical-infrared countermeasure
EOL	end of link
EOM	end of message
EOP	emergency operating procedures
E-O TDA	electro-optical tactical decision aid
EOW	engineering orderwire
EP	electronic protection; emergency preparedness; emergency procedures; execution planning
EPA	Environmental Protection Agency; evasion plan of action
EPBX	electronic private branch exchange
EPF	enhanced palletized load system (PLS) flatrack
EPH	emergency planning handbook
EPIC	El Paso Intelligence Center
EPIRB	emergency position-indicating radio beacon
EPLO	emergency preparedness liaison officer
EPROM	erasable programmable read-only memory
EPW	enemy prisoner of war
EPW/CI	enemy prisoner of war/civilian internee
ERC	exercise related construction
ERDC	Engineer Research and Development Center
ERGM	extended range guided munitions
ERO	engine running on or offload
ERRO	Emergency Response and Recovery Office
ERSD	estimated return to service date
ERT	emergency response team (FEMA); engineer reconnaissance team
ERT-A	emergency response team - advance element
ES	electronic warfare support
ESB	engineer support battalion
ESC	Electronics Systems Center
ESF	Economic Support Fund; emergency support function
ESG	executive steering group; expeditionary strike group
ESGN	electrically suspended gyro navigation
ESI	extremely sensitive information
ESK	electronic staff weather officer kit
ESM	expeditionary site mapping
ESO	embarkation staff officer; environmental science officer
ESOC	Emergency Supply Operations Center
ESP	engineer support plan
ESR	external supported recovery

EST	embarked security team; emergency service team; emergency support team (FEMA); en route support team
ETA	estimated time of arrival
ETAC	emergency tactical air control; enlisted terminal attack controller
ETD	estimated time of departure
ETF	electronic target folder
ETI	estimated time of intercept
ETIC	estimated time for completion; estimated time in commission
ETM	electronic transmission
ETPL	endorsed TEMPEST products list
ETR	export traffic release
ETS	European telephone system
ETSS	extended training service specialist
ETX	end of text
EU	European Union
E-UAV	endurance unmanned aerial vehicle
EUB	essential user bypass
EURV	essential user rekeying variable
EUSA	Eighth US Army
EUSC	effective United States control/controlled
EUSCS	effective United States controlled ships
EVC	evasion chart
EVE	equal value exchange
EW	early warning; electronic warfare
EWC	electronic warfare coordinator
EWCC	electronic warfare coordination cell
EWCS	electronic warfare control ship
EW/GCI	early warning/ground-controlled intercept
EWIR	electronic warfare integrated reprogramming
EWO	electronic warfare officer
EXCIMS	Executive Council for Modeling and Simulations
EXCOM	extended communications search
ExCom	executive committee
EXDIR	Executive Director (CIA)
EXDIR/ICA	Executive Director for Intelligence Community Affairs (USG)
EXECSEC	executive secretary
EXER	exercise
EXORD	execute order
EXPLAN	exercise plan
EZ	extraction zone
EZCO	extraction zone control officer
EZM	engagement zone manager

F	flare patterns; flash
F2T2EA	find, fix, track, target, engage, and assess
F&ES	fire and emergency services
FA	feasibility assessment; field artillery
FAA	Federal Aviation Administration; Foreign Assistance Act
FAAR	facilitated after-action review
FAC	forward air controller

FAC(A)	forward air controller (airborne)
FACE	forward aviation combat engineering
FACSFAC	fleet area control and surveillance facility
FAD	feasible arrival date; force activity designator
FAE	fuel air explosive
FALD	Field Administration and Logistics Division
FAM	functional area manager
FAMP	forward area minefield planning
FAO	Food and Agriculture Organization (UN) foreign area officer
FAPES	Force Augmentation Planning and Execution System
FAR	Federal Acquisition Regulation; Federal Aviation Regulation
FARC	Revolutionary Armed Forces of Colombia
FARP	forward arming and refueling point
FAS	Foreign Agricultural Service (USDA); Frequency assignment subcommittee; fueling at sea; functional account symbol
FASCAM	family of scatterable mines
FAST	fleet antiterrorism security team
FAX	facsimile
FB	forward boundary
FBI	Federal Bureau of Investigation
FBIS	Foreign Broadcast Information Service
FBO	faith-based organization
FC	field circular; final coordination; floating causeway; floating craft; force commander
FCA	Foreign Claims Act; functional configuration audit
FCC	Federal Communications Commission; Federal coordinating center; functional combatant commander
FCE	forward command element
FCG	foreign clearance guide
FCM	foreign consequence management
FCO	Federal coordinating officer
FCP	fire control party
FCT	firepower control team
FD	from temporary duty
FDA	Food and Drug Administration
FDBM	functional database manager
FDC	fire direction center
FDESC	force description
FDL	fast deployment logistics
FDLP	flight deck landing practice
FDM	frequency division multiplexing
FDO	fire direction officer; flexible deterrent option; flight deck officer; foreign disclosure officer
FDR/FA	flight data recorder/fault analyzer
FDS	fault detection system
FDSL	fixed directory subscriber list
FDSS	fault detection subsystem
FDSSS	flight deck status and signaling system
FDT	forward distribution team
FDUL	fixed directory unit list

FDX	full duplex
FE	facilities engineering
FEA	front-end analysis
FEBA	forward edge of the battle area
FEC	forward error correction
FECC	fires and effects coordination cell
FED-STD	federal standard
FEK	frequency exchange keying
FEMA	Federal Emergency Management Agency
FEP	fleet satellite (FLTSAT) extremely high frequency (EHF) package
FEPP	federal excess personal property; foreign excess personal property
FEST	foreign emergency support team; forward engineer support team
FET	facility engineer team
FEU	forty-foot equivalent unit
FEZ	fighter engagement zone
FF	navy fast frigate
Ff	fatigue correction factor
FFA	free-fire area
FFC	force fires coordinator
FFCC	flight ferry control center; force fires coordination center
FFD	foundation feature data
FFE	field force engineering; fire for effect; flame field expedients
FFG	guided missile frigate
FFH	fast frequency hopping
FFH-net	fast-frequency-hopping net
FFHT-net	fast-frequency-hopping training net
FFIR	friendly force information requirement
FFP	fresh frozen plasma
FFTU	forward freight terminal unit
FG	fighter group
FGMDSS	Future Global Maritime Distress and Safety System
FGS	final governing standard
FH	fleet hospital
FHA	Bureau for Food and Humanitarian Assistance; Federal Highway Administration; foreign humanitarian assistance
FHC	family help center
F-hour	effective time of announcement by the Secretary of Defense to the Military Departments of a decision to mobilize Reserve units
FHP	force health protection
FHWA	Federal Highway Administration
FI	foreign intelligence
FIA	functional interoperability architecture
FIC	force indicator code
FID	foreign internal defense
FIDAF	foreign internal defense augmentation force
FIE	fly-in echelon
FIFO	first-in-first-out
FIR	first-impressions report; flight information region
FIRCAP	foreign intelligence requirements capabilities and priorities
1st IOC	1st Information Operations Command (Land)

FIS	flight information service; Foreign Intelligence Service
FISC	fleet and industrial supply center
FISINT	foreign instrumentation signals intelligence
FISS	foreign intelligence and security services
FIST	fire support team; fleet imagery support terminal; fleet intelligence support team
FIWC	fleet information warfare center
FIXe	navigational fix error
FLAR	forward-looking airborne radar
FLENUMMETOCCEN	Fleet Numerical Meteorology and Oceanography Center
FLENUMMETOCDET	Fleet Numerical Meteorological and Oceanographic Detachment
FLETC	Federal Law Enforcement Training Center
FLIP	flight information publication; flight instruction procedures
FLIR	forward-looking infrared
FLITE	federal legal information through electronics
FLO/FLO	float-on/float-off
FLOLS	fresnel lens optical landing system
FLOT	forward line of own troops
FLP	force level planning
FLS	forward logistic site
FLSG	force logistic support group
FLTSAT	fleet satellite
FLTSATCOM	fleet satellite communications
FM	field manual (Army); financial management; flare multiunit; force module; frequency modulation; functional manager
FMA-net	frequency management A-net
FMAS	foreign media analysis subsystem
FMAT	financial management augmentation team
FMC	force movement characteristics; full mission-capable
FMCH	fleet multichannel
FMCR	Fleet Marine Corps Reserve
FMCSA	Federal Motor Carrier Safety Administration
FMI	field manual-interim
FMF	Fleet Marine Force
FMFP	foreign military financing program
FMID	force module identifier
FMO	frequency management office
FMP	force module package; foreign materiel program
FMS	force module subsystem; foreign military sales
FMSC	frequency management sub-committee
FMT-net	frequency management training net
FN	foreign nation
FNMOC	Fleet Numerical Meteorology and Oceanographic Center
FNMOD	Fleet Numerical Meteorological and Oceanographic Detachment
FNOC	Fleet Numerical Oceanographic Command
FNS	foreign nation support
FO	fiber optic; flash override; forward observer
FOB	forward operating base; forward operations base
FOC	full operational capability; future operations cell
FOD	field operations division; foreign object damage

FOFW	fiber optic field wire
FOI	fault detection isolation
FOIA	Freedom of Information Act
FOIU	fiber optic interface unit
FOL	fiber optic link; forward operating location
FON	freedom of navigation (operations)
FORSCOM	United States Army Forces Command
FORSTAT	force status and identity report
FOS	forward operating site; full operational status
FOT	follow-on operational test
FOUO	for official use only
FOV	field of view
FP	firing point force protection; frequency panel
FPA	foreign policy advisor
FPC	field press censorship; final planning conference; future plans cell
FPCON	force protection condition
FPD	force protection detachment; foreign post differential
FPF	final protective fire
FPM	Federal personnel manual
FPO	Fleet post office
FPOC	focal point operations center
FPS	force protection source
FPTAS	flight path threat analysis simulation
FPTS	forward propagation by tropospheric scatter
FPWG	force protection working group
FR	final report; frequency response
FRA	Federal Railroad Administration (DOT)
FRAG	fragmentation code
FRAGORD	fragmentary order
FRC	federal resource coordinator; forward resuscitative care
FRD	formerly restricted data
FREQ	frequency
FRERP	Federal Radiological Emergency Response Plan
FRF	fragment retention film
FRMAC	Federal Radiological Monitoring and Assessment Center (DOE)
FRN	force requirement number
FROG	free rocket over ground
FRP	Federal response plan (USG)
FRRS	frequency resource record system
FS	fighter squadron; file separator; file server; flare single-unit
fs	search radius safety factor
FSA	fire support area
FSB	fire support base; forward staging base; forward support base; forward support battalion
FSC	fire support cell; fire support coordinator
FSCC	fire support coordination center
FSCL	fire support coordination line
FSCM	fire support coordination measure
FSCOORD	fire support coordinator
FSE	fire support element

FSEM	fire support execution matrix
FSK	frequency shift key
FSN	foreign service national
FSO	fire support officer; flight safety officer; foreign service officer
FSS	fast sealift ships; fire support station; flight service station
FSSG	force service support group (USMC)
FSST	forward space support to theater
FST	fleet surgical team
FSU	former Soviet Union; forward support unit
FSW	feet of seawater
ft	feet; foot
ft3	cubic feet
FTC	Federal Trade Commission
FTCA	Foreign Tort Claims Act
FTP	file transfer protocol
FTRG	fleet tactical readiness group
FTS	Federal Telecommunications System; Federal telephone service; file transfer service
FTU	field training unit; freight terminal unit
FTX	field training exercise
FUAC	functional area code
FUNCPLAN	functional plan
F/V	fishing vessel
Fv	aircraft speed correction factor
FVT	Force Validation Tool
FW	fighter wing; weather correction factor
FWD	forward
FWDA	friendly weapon danger area
FWF	former warring factions
FY	fiscal year
FYDP	Future Years Defense Program
G-1	Army or Marine Corps component manpower or personnel staff officer (Army division or higher staff, Marine Corps brigade or higher staff)
G-2	Army or Marine Corps component intelligence staff officer(Army division or higher staff, Marine Corps brigade or higher staff)
G-3	Army or Marine Corps component operations staff officer (Army division or higher staff, Marine Corps brigade or higher staff)
G-4	Army or Marine Corps component logistics staff officer (Army division or higher staff, Marine Corps brigade or higher staff); Assistant Chief of Staff for Logistics
G-6	Army or Marine Corps component command, control, communications, and computer systems staff officer
G-7	information operations staff officer (ARFOR)
G/A	ground to air
GA	Tabun, a nerve agent
GAA	general agency agreement; geospatial intelligence assessment activity
GAFS	General Accounting and Finance System
GAMSS	Global Air Mobility Support System
GAO	General Accounting Office; Government Accountability Office

GAR	gateway access request
GARS	Global Area Reference System
GATB	guidance, apportionment, and targeting board
GATES	Global Air Transportation Execution System
GB	group buffer; Sarin, a nerve agent
GBL	government bill of lading
GBR	ground-based radar
GBS	Global Broadcast Service; Global Broadcast System
GBU	guided bomb unit
GC	general counsel; Geneva Convention; Geneva Convention Relative to the Protection of Civilian Persons in Time of War
GC3A	global command, control, and communications assessment
GC4A	global command, control, communications, and computer assessment
GCA	ground controlled approach
GCC	geographic combatant commander; global contingency construction
GCCS	Global Command and Control System
GCCS-A	Global Command and Control System-Army
GCCS-I3	Global Command and Control System Integrated Imagery and Intelligence
GCCS-J	Global Command and Control System-Joint
GCCS-M	Global Command and Control System-Maritime
GCE	ground combat element (MAGTF)
GCI	ground control intercept
GCP	geospatial-intelligence contingency package; Ground commander's pointer
GCRI	general collective routing indicator (RI)
GCS	ground control station
GCSS	Global Combat Support System
GD	Soman, a nerve agent
GDF	gridded data field
GDIP	General Defense Intelligence Program
GDIPP	General Defense Intelligence Proposed Program
GDP	General Defense Plan (SACEUR): gross domestic product
GDSS	Global Decision Support System
GE	general engineering
GEM	Global Information Grid (GIG) Enterprise Management
GENADMIN	general admin (message)
GENSER	general service (message)
GENTEXT	general text
GEO	geosynchronous earth orbit
GEOCODE	geographic code
GEOFILE	geolocation code file; standard specified geographic location file
GEOINT	geospatial intelligence
GEOLOC	geographic location; geographic location code
GEOREF	geographic reference; world geographic reference system
GF	a nerve agent
GFE	government-furnished equipment
GFI	government-furnished information
GFM	Global Force Management; global freight management; government-furnished material

GFMPL	Graphics Fleet Mission Program Library
GFOAR	global family of operation plans assessment report
GFU	group framing unit
GHz	gigahertz
GI	geomatics and imagery
GI&S	geospatial information and services
GIAC	graphic input aggregate control
GIC	(*gabarit international de chargement)* international loading gauge
GIE	global information environment
GIG	Global Information Grid
GII	global information infrastructure
GIP	gridded installation photograph
GIS	geographic information system; geospatial information systems
GLCM	ground launched cruise missile
GLINT	gated laser intensifier
GLO	ground liaison officer
GLTD	ground laser target designator
GM	group modem
GMD	global missile defense; group mux and/or demux
GMDSS	Global Maritime Distress and Safety System
GMF	ground mobile force
GMFP	global military force policy
GMI	general military intelligence
GMR	graduated mobilization response; ground mobile radar
GMT	Greenwich Mean Time
GMTI	ground moving target indicator
GNC	Global Network Operations Center
GND	Global Information Grid (GIG) Network Defense
GOCO	government-owned, contractor-operated
GOES	geostationary operational environmental satellite
GOGO	government-owned, government-operated
GOS	grade of service
GOSG	general officer steering group
GOTS	government off-the-shelf
GP	general purpose; group
GPC	government purchase card
GPD	gallons per day
GPE	geospatial intelligence preparation of the environment
GPEE	general purpose encryption equipment
GPL	Geospatial Product Library
GPM	gallons per minute
GPMDM	group modem
GPMRC	Global Patient Movement Requirements Center
GPS	global positioning system
GPW	Geneva Convention Relative to the Treatment of Prisoners of War
GR	graduated response
GRASP	general retrieval and sort processor
GRCA	ground reference coverage area
GRG	gridded reference graphic
GRL	global reach laydown

GRREG	graves registration
GS	general service; general support; ground speed; group separator
GSA	General Services Administration; general support artillery
GSE	ground support equipment
GSI	glide slope indicator
GSM	ground station module
GSO	general services officer
GSORTS	Global Status of Resources and Training System
GS-R	general support-reinforcing
GSR	general support-reinforcing; ground surveillance radar
GSSA	general supply support area
gt	gross ton
GTAS	ground-to-air signals
GTL	gun-target line
GTM	global transportation management
GTN	Global Transportation Network
GUARD	US National Guard and Air Guard
GUARDS	General Unified Ammunition Reporting Data System
G/VLLD	ground/vehicle laser locator designator
GW	guerrilla warfare
GWC	global weather central
GWEN	Ground Wave Emergency Network
GWOT	global war on terror
GWS	Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field
GWS Sea	Geneva Convention for the Amelioration of the Condition of the Wounded, Sick, and Shipwrecked Members of the Armed Forces at Sea
H&I	harassing and interdicting
H&S	headquarters and service
HA	holding area; humanitarian assistance
HAARS	high-altitude airdrop resupply system
HAB	high altitude burst
HAC	helicopter aircraft commander
HACC	humanitarian assistance coordination center
HAHO	high-altitude high-opening parachute technique
HALO	high-altitude low-opening parachute technique
HAP	humanitarian assistance program
HAP-EP	humanitarian assistance program-excess property
HARM	high-speed antiradiation missile
HARP	high altitude release point
HAST	humanitarian assistance survey team
HATR	hazardous air traffic report
HAZ	hazardous cargo
HAZMAT	hazardous materials
HB	heavy boat
HBCT	heavy brigade combat team
HCA	head of contracting authority; humanitarian and civic assistance
HCAS	hostile casualty
HCL	hydrochloride

HCO	helicopter control officer
HCP	hardcopy printer
HCS	helicopter combat support (Navy) helicopter control station; helicopter coordination section
HCT	human intelligence (HUMINT) collection team
HD	a mustard agent; harmonic distortion; homeland defense
HDC	harbor defense commander; helicopter direction center
HDCU	harbor defense command unit
HDM	humanitarian demining
HDO	humanitarian demining operations
HDPLX	half duplex
HE	heavy equipment; high explosive
HEAT	helicopter external air transport; high explosive antitank
HEC	helicopter element coordinator
HEFOE	hydraulic electrical fuel oxygen engine
HEI	high explosives incendiary
HEL-H	heavy helicopter
HEL-L	light helicopter
HEL-M	medium helicopter
HELO	helicopter
HEMP	high-altitude electromagnetic pulse
HEMTT	heavy expanded mobile tactical truck
HEO	highly elliptical orbit
HEPA	high efficiency particulate air
HERF	hazards of electromagnetic radiation to fuel
HERO	electromagnetic radiation hazards; hazards of electromagnetic radiation to ordnance
HERP	hazards of electromagnetic radiation to personnel
HET	heavy equipment transporter; human intelligence (HUMINT) exploitation team
HF	high frequency
HFDF	high frequency direction-finding
HFRB	high frequency regional broadcast
HH	homing pattern
HHD	headquarters and headquarters detachment
H-hour	seaborne assault landing hour; specific time an operation or exercise begins
HHQ	higher headquarters
HHS	Department of Health and Human Services
HICAP	high-capacity firefighting foam station
HIDACZ	high-density airspace control zone
HIDTA	high-intensity drug trafficking area
HIFR	helicopter in-flight refueling
HIMAD	high to medium altitude air defense
HIMARS	High Mobility Artillery Rocket System
HIMEZ	high-altitude missile engagement zone
HIRSS	hover infrared suppressor subsystem
HIV	human immuno-deficiency virus
HJ	crypto key change
HLPS	heavy-lift pre-position ship

HLZ	helicopter landing zone
HM	hazardous material
HMA	humanitarian mine action
HMH	Marine heavy helicopter squadron
HMIS	Hazardous Material Information System
HMLA	Marine light/attack helicopter squadron
HMM	Marine medium helicopter squadron
HMMWV	high mobility multipurpose wheeled vehicle (Humvee, Hummer)
HMOD	harbormaster operations detachment
HMW	health, morale, and welfare
HN	host nation
HNS	host-nation support
HNSA	host-nation support agreement
HNSCC	host-nation support coordination cell
HOB	height of burst
HOC	human intelligence operations cell; Humanitarian operations center
HOCC	humanitarian operations coordination center
HOD	head of delegation
HOGE	hover out of ground effect
HOIS	hostile intelligence service
HOM	head of mission
HOSTAC	helicopter operations from ships other than aircraft carriers (USN publication)
HOTPHOTOREP	hot photo interpretation report
HPA	high power amplifier
HPMSK	high priority mission support kit
HPT	high-payoff target
HPTL	high-payoff target list
HQ	headquarters; have quick
HQCOMDT	headquarters commandant
HQDA	Headquarters, Department of the Army
HQFM-net	HAVE QUICK frequency modulation net
HQFMT-net	HAVE QUICK frequency modulation training net
HQMC	Headquarters, Marine Corps
HR	helicopter request; hostage rescue
HRB	high-risk billet
HRC	high-risk-of-capture
HRJTF	humanitarian relief joint task force
HRO	humanitarian relief organizations
HRP	high-risk personnel; human remains pouch
HRS	horizon reference system
HRT	hostage rescue team
HS	helicopter antisubmarine (Navy); homeland security; homing single-unit
HSAC	Homeland Security Advisory Council
HSAS	Homeland Security Advisory System
HSB	high speed boat
HSC	helicopter sea combat (Navy); Homeland Security Council
HSCDM	high speed cable driver modem
HSC/PC	Homeland Security Council Principals Committee

HSC/PCC	Homeland Security Council Policy Coordination Committee
HSD	human intelligence support detachment
HSE	headquarters support element; human intelligence support element (DIA)
HSEP	hospital surgical expansion package (USAF)
HSI	hyperspectral imagery
HSLS	health service logistic support
HSM	humanitarian service medal
HSPD	homeland security Presidential directive
HSPR	high speed pulse restorer
HSS	health service support
HSSDB	high speed serial data buffer
HST	helicopter support team
HT	hatch team
HTERRCAS	hostile terrorist casualty
HTG	hard target graphic
HTH	high test hypochlorite
HU	hospital unit
HUD	head-up display
HUMINT	human intelligence; human resources intelligence
HUMRO	humanitarian relief operation
HUMVEE	High Mobility Multipurpose Wheeled vehicle (HMMWV), Hummer
HUS	hardened unique storage
HVA	high-value asset
HVAA	high-value airborne asset
HVAC	heating, ventilation, and air conditioning
HVI	high-value individual
HVT	high-value target
HW	hazardous waste
HWM	high water mark
Hz	hertz
I	immediate; individual
I&A	Office of Intelligence and Analysis
I&W	indications and warning
IA	implementing arrangement; individual augmentee; information assurance; initial assessment
IAC	Interagency Advisory Council
IACG	interagency coordination group
IADB	Inter-American Defense Board
IADS	integrated air defense system
IAEA	International Atomic Energy Agency (UN)
IAF	initial approach fix
IAIP	Information Analysis and Infrastructure Protection
IAMSAR	International Aeronautical and Maritime Search and Rescue manual
IAP	international airport
IAR	interoperability assessment report
IASC	Interagency Standing Committee (UN); interim acting service chief
IATA	International Air Transport Association
IATACS	Improved Army Tactical Communications System

IATO	interim authority to operate
IAVM	information assurance vulnerability management
IAW	in accordance with
I/B	inboard
IBB	International Broadcasting Bureau
IBCT	infantry brigade combat team
IBES	intelligence budget estimate submission
IBM	International Business Machines
IBS	Integrated Booking System; integrated broadcast service; Integrated Broadcast System
IBU	inshore boat unit
IC	incident commander; intelligence community; intercept
IC3	integrated command, control, and communications
ICAD	individual concern and deficiency
ICAO	International Civil Aviation Organization
ICBM	intercontinental ballistic missile
ICC	information coordination center; Intelligence Coordination Center; Interstate Commerce Commission
ICD	international classifications of diseases; International Cooperation and Development Program (USDA)
ICDC	Intelligence Community Deputies Committee
ICE	Immigration and Customs Enforcement
ICEDEFOR	Iceland Defense Forces
IC/EXCOM	Intelligence Community Executive Committee
ICF	intelligence contingency funds
ICG	interagency core group
ICIS	integrated consumable item support
ICITAP	International Crime Investigative Training Assistance Program (DOJ)
ICM	image city map; improved conventional munitions; Integrated collection management
ICN	idle channel noise; interface control net
ICNIA	integrated communications, navigation, and identification avionics
ICOD	intelligence cutoff data
ICODES	integrated computerized deployment system
ICON	imagery communications and operations node; intermediate coordination node
ICP	incident command post; intertheater communications security (COMSEC) package; interface change proposal; inventory control point
ICPC	Intelligence Community Principals Committee
ICR	Intelligence Collection Requirements
ICRC	International Committee of the Red Cross
ICRI	interswitch collective routing indicator
ICS	incident command system; internal communications system; inter-Service chaplain support
ICSF	integrated command communications system framework
ICSAR	interagency committee on search and rescue
ICU	intensive care unit; interface control unit
ICVA	International Council of Voluntary Agencies
ICW	in coordination with
ID	identification; initiating directive

IDAD	internal defense and development
IDB	integrated database
IDCA	International Development Cooperation Agency
IDDF	intermediate data distribution facility
IDEAS	Intelligence Data Elements Authorized Standards
IDEX	imagery data exploitation system
IDF	intermediate distribution frame
IDHS	intelligence data handling system
IDM	improved data modem; information dissemination management
IDNDR	International Decade for Natural Disaster Reduction (UN)
IDO	installation deployment officer
IDP	imagery derived product; imminent danger pay; Internally displaced person
IDRA	infectious disease risk assessment
IDS	individual deployment site; integrated deployment system; interface design standards; intrusion detection system
IDSS	interoperability decision support system
IDT	inactive duty training
IDZ	inner defense zone
IEB	intelligence exploitation base
IED	improvised explosive device
IEDD	improvised explosive device defeat
IEEE	Institute of Electrical and Electronics Engineers
IEL	illustrative evaluation scenario
IEMATS	improved emergency message automatic transmission system
IER	information exchange requirement
IES	imagery exploitation system
IESS	imagery exploitation support system
IEW	intelligence and electronic warfare
IF	intermediate frequency
IFC	intelligence fusion center
IFCS	improved fire control system
IFF	identification, friend or foe
IFFN	identification, friend, foe, or neutral
IFF/SIF	identification, friend or foe/selective identification feature
IFP	integrated force package
IFR	instrument flight rules
IFRC	International Federation of Red Cross and Red Crescent Societies
IFSAR	interferometric synthetic aperture radar
IG	inspector general
IGL	intelligence gain/loss
IGO	intergovernmental organization
IGSM	interim ground station module (JSTARS)
IHADSS	integrated helmet and display sight system (Army)
IHC	International Humanitarian Community
IHO	industrial hygiene officer
IHS	international health specialist
IIB	interagency information bureau
IICL	Institute of International Container Lessors
III	incapacitating illness or injury

IIM	intelligence information management
IIP	international information program; Interoperability improvement program
IIR	imagery interpretation report; imaging infrared; intelligence information report
IJC3S	initial joint command, control, and communications system; Integrated Joint Command, Control, and Communications System
IL	intermediate location
ILO	International Labor Organization (UN)
ILOC	integrated line of communications
ILS	instrument landing system; integrated logistic support
IM	information management
IMA	individual mobilization augmentee
IMC	instrument meteorological conditions; International Medical Corps
IMDC	isolated, missing, detained, or captured
IMDG	international maritime dangerous goods (UN)
IMET	international military education and training
IMETS	Integrated Meteorological System
IMF	International Monetary Fund (UN)
IMI	international military information
IMINT	imagery intelligence
IMIT	international military information team
IMLTU	intermatrix line termination unit
IMM	integrated materiel management
IMMDELREQ	immediate delivery required
IMO	information management officer; International Maritime Organization
IMOSAR	International Maritime Organization (IMO) search and rescue manual
IMOSS	interim mobile oceanographic support system
IMP	implementation; information management plan; inventory management plan
IMPT	incident management planning team
IMS	information management system; international military staff; international military standardization
IMSU	installation medical support unit
IMU	inertial measuring unit; intermatrix unit IN Air Force component intelligence officer (staff); impulse noise; instructor
INCERFA	uncertainty phase (ICAO)
INCNR	increment number
INCSEA	incidents at sea
IND	improvised nuclear device
INF	infantry
INFLTREP	inflight report
INFOCON	information operations condition
INFOSEC	information security
ING	Inactive National Guard
INID	intercept network in dialing
INJILL	injured or ill
INL	Bureau for International Narcotics and Law Enforcement Affairs (USG)
INM	international narcotics matters
INMARSAT	international maritime satellite

INR	Bureau of Intelligence and Research, Department of State
INREQ	information request
INRP	Initial National Response Plan
INS	Immigration and Naturalization Service; inertial navigation system; insert code
INSCOM	United States Army Intelligence and Security Command
INTAC	individual terrorism awareness course
INTACS	integrated tactical communications system
INTELSAT	International Telecommunications Satellite Organization
INTELSITSUM	intelligence situation summary
InterAction	American Council for Voluntary International Action
INTERCO	International Code of signals
INTERPOL	International Criminal Police Organization
INTERPOL-USNCB	International Criminal Police Organization, United States National Central Bureau (DOJ)
INTREP	intelligence report
INTSUM	intelligence summary
INU	inertial navigation unit; integration unit
INV	invalid
INVOL	involuntary
I/O	input/output
IO	information objectives; information operations; intelligence oversight
IOC	Industrial Operations Command; initial operational capability; intelligence operations center; investigations operations center
IOI	injured other than hostilities or illness
IOM	installation, operation, and maintenance; International Organization for Migration
IOP	interface operating procedure
IOSS	Interagency Operations Security (OPSEC) Support Staff
IOU	input/output unit
IOWG	information operations working group
IP	initial point; initial position; instructor pilot; internet protocol
IPA	intelligence production agency
IPB	intelligence preparation of the battlespace
IPBD	intelligence program budget decision
IPC	initial planning conference; interagency planning cell
IPDM	intelligence program decision memorandum
IPDP	inland petroleum distribution plan
IPDS	imagery processing and dissemination system; Inland petroleum distribution system (Army)
IPE	individual protective equipment; industrial plant equipment
IPG	isolated personnel guidance
IPI	indigenous populations and institutions
IPIR	initial photo interpretation report
IPL	imagery product library; integrated priority list
IPO	International Program Office
IPOE	intelligence preparation of the operational environment
IPOM	intelligence program objective memorandum
IPP	impact point prediction; industrial preparedness program
IPR	in-progress review; intelligence production requirement

IPRG	intelligence program review group
IPS	illustrative planning scenario; Interim Polar System; interoperability planning system
IPSG	intelligence program support group
IPSP	intelligence priorities for strategic planning
IPT	integrated planning team integrated process team; Integrated Product Team
I/R	internment/resettlement
IR	incident report; information rate; information requirement; infrared; intelligence requirement
IRAC	interdepartment radio advisory committee
IRC	International Red Cross; International Rescue Committee
IRCCM	infrared counter countermeasures
IRCM	infrared countermeasures
IRDS	infrared detection set
IRF	Immediate Reaction Forces (NATO); incident response force
IRINT	infrared intelligence
IRISA	Intelligence Report Index Summary File
IRO	international relief organization
IR pointer	infrared pointer
IRR	Individual Ready Reserve; integrated readiness report
IRS	Internal Revenue Service
IRST	infrared search and track
IRSTS	infrared search and track sensor; Infrared Search and Track System
IRT	Initial Response Team
IS	information superiority; information system interswitch
ISA	international standardization agreement; inter-Service agreement
ISAF	International Security Assistance Force
ISB	intermediate staging base
ISDB	integrated satellite communications (SATCOM) database
ISE	intelligence support element
ISG	information synchronization group
ISMCS	international station meteorological climatic summary
ISMMP	integrated continental United States (CONUS) medical mobilization plan
ISO	International Organization for Standardization; isolation
ISOO	Information Security Oversight Office
ISOPAK	International Organization for Standardization package
ISOPREP	isolated personnel report
ISP	internet service provider
ISR	intelligence, surveillance, and reconnaissance
ISS	in-system select
ISSA	inter-Service support agreement
ISSG	Intelligence Senior Steering Group
ISSM	information system security manager
ISSO	information systems security organization
IST	integrated system test; interswitch trunk
IT	information technology
ITA	international telegraphic alphabet
ITAC	intelligence and threat analysis center (Army)

ITALD	improved tactical air-launched decoy
ITAR	international traffic in arms regulation (coassembly)
ITF	intelligence task force (DIA)
ITG	infrared target graphic
ITO	installation transportation officer
ITRO	inter-Service training organization
ITU	International Telecommunications Union
ITV	in-transit visibility
ITW/AA	integrated tactical warning and attack assessment
IUWG	inshore undersea warfare group
IV	intravenous
IVR	initial voice report
IVSN	Initial Voice Switched Network
IW	irregular warfare
IWC	information operations warfare commander
IW-D	defensive information warfare
IWG	intelligence working group; interagency working group
IWSC	Information Warfare Support Center
IWW	inland waterway
IWWS	inland waterway system
J-1	manpower and personnel directorate of a joint staff; manpower and personnel staff section
J-2	intelligence directorate of a joint staff; intelligence staff section
J-2A	deputy directorate for administration of a joint staff
J2-CI	Joint Counterintelligence Office
J-2J	deputy directorate for support of a joint staff
J-2M	deputy directorate for crisis management of a joint staff
J-2O	deputy directorate for crisis operations of a joint staff
J-2P	deputy directorate for assessment, doctrine, requirements, and capabilities of a joint staff
J-2T	Deputy Directorate for Targeting, Joint Staff Intelligence Directorate
J-2T-1	joint staff target operations division
J-2T-2	Target Plans Division
J-2X	joint force staff counterintelligence and human intelligence element
J-3	operations directorate of a joint staff; operations staff section
J-4	logistics directorate of a joint staff; logistics staff section
J-5	plans directorate of a joint staff; plans staff section
J-6	communications system directorate of a joint staff; command, control, communications, and computer systems staff section
J-7	engineering staff section; perational plans and interoperability directorate of a joint staff
J-7/JED	exercises and training directorate of a joint staff
J-8	Director for Force Structure, Resource, and Assessment, Joint Staff; force structure, resource, and assessment directorate of a joint staff
J-9	civil-military operations staff section
JA	judge advocate
J-A	judge advocate directorate of a joint staff
JAAP	joint airborne advance party
JAAR	joint after-action report

JAARS	Joint After-Action Reporting System
JAAT	joint air attack team
JA/ATT	joint airborne and air transportability training
JAC	joint analysis center
JACC	joint airspace control center
JACCC	joint airlift coordination and control cell
JACC/CP	joint airborne communications center/command post
JACE	joint air coordination element
JACS	joint automated communication-electronics operating instructions system
JADO	joint air defense operations
JADOCS	Joint Automated Deep Operations Coordination System
JAFWIN	JWICS Air Force weather information network
JAG	Judge Advocate General
JAGMAN	Manual of the Judge Advocate General (US Navy)
JAI	joint administrative instruction; joint airdrop inspection
JAIC	joint air intelligence center
JAIEG	joint atomic information exchange group
JAMPS	Joint Interoperability of Tactical Command and Control Systems (JINTACCS) automated message preparation system
JANAP	Joint Army, Navy, Air Force publication
JAO	joint air operations
JAOC	joint air operations center
JAOP	joint air operations plan
JAPO	joint area petroleum office
JAR	joint activity report
JARB	joint acquisition review board
JARCC	joint air reconnaissance control center
JARS	joint automated readiness system
JASC	joint action steering committee
JASSM	Joint Air-to-Surface Standoff Missile
JAT	joint acceptance test
JATACS	joint advanced tactical cryptological support
JATF	joint amphibious task force
JAT Guide	Joint Antiterrorism Program Manager's Guide
JAWS	Joint Munitions Effectiveness Manual (JMEM)/air-to-surface weaponeering system
JBP	Joint Blood Program
JBPO	joint blood program office
JC2WC	joint command and control warfare center
JCA	jamming control authority; Joint Capability Area
JCASREP	joint casualty report
JCAT	joint crisis action team
JCC	joint command center; joint contracting center; joint course catalog
JCCB	Joint Configuration Control Board
JCCC	joint combat camera center
JCCP	joint casualty collection point
JCE	Joint Intelligence Virtual Architecture (JIVA) Collaborative Environment
JCEOI	joint communications-electronics operating instructions

JCET	joint combined exchange training; joint combined exercise for training
JCEWR	joint coordination of electronic warfare reprogramming
JCEWS	joint force commander's electronic warfare staff
JCGRO	joint central graves registration office
JCIDO	Joint Combat Identification Office
JCIOC	joint counterintelligence operations center
JCISA	Joint Command Information Systems Activity
JCISB	Joint Counterintelligence Support Branch
JCLL	joint center for lessons learned
JCMA	joint communications security monitor activity
JCMB	Joint Collection Management Board
JCMC	joint crisis management capability
JCMEB	joint civil-military engineering board
JCMEC	joint captured materiel exploitation center
JCMO	joint communications security management office
JCMOTF	joint civil-military operations task force
JCMPO	Joint Cruise Missile Project Office
JCMT	joint collection management tools
JCN	joint communications network
JCS	Joint Chiefs of Staff
JCSAN	Joint Chiefs of Staff Alerting Network
JCSAR	joint combat search and rescue
JCSC	joint communications satellite center
JCSE	joint communications support element
JCSM	Joint Chiefs of Staff memorandum
JCSP	joint contracting support plan
JCSS	joint communications support squadron
JCTN	joint composite track network
JDA	joint duty assignment
JDAAP	Joint Doctrine Awareness Action Plan
JDAL	Joint Duty Assignment List
JDAM	Joint Direct Attack Munition
JDAMIS	Joint Duty Assignment Management Information System
JDC	joint deployment community; Joint Doctrine Center
JDD	joint doctrine distribution
JDDC	joint doctrine development community
JDDOC	joint deployment distribution operations center
JDEC	joint document exploitation center
JDEIS	Joint Doctrine, Education, and Training Electronic Information System
JDIG	Joint Drug Intelligence Group
JDISS	joint deployable intelligence support system
JDN	joint data network
JDNO	joint data network operations officer
JDOG	joint detention operations group
JDOMS	Joint Director of Military Support
JDPC	Joint Doctrine Planning Conference
JDPO	joint deployment process owner
JDSS	Joint Decision Support System
JDSSC	Joint Data Systems Support Center
JDST	joint decision support tools

JDTC	Joint Deployment Training Center
JE	joint experimentation
JEAP	Joint Electronic Intelligence (ELINT) Analysis Program
JECG	joint exercise control group
JECPO	Joint Electronic Commerce Program Office
JEDD	Joint Education and Doctrine Division
JEEP	joint emergency evacuation plan
JEL	Joint Electronic Library
JEM	joint exercise manual
JEMB	joint environmental management board
JEMP	joint exercise management package
JEPES	joint engineer planning and execution system
JET	Joint Operation Planning and Execution System (JOPES) editing tool
JEWC	Joint Electronic Warfare Center
JEZ	joint engagement zone
JFA	joint field activity
JFACC	joint force air component commander
JFAST	Joint Flow and Analysis System for Transportation
JFC	joint force commander
JFCC	joint functional component command
JFCC-IMD	Joint Functional Component Command for Integrated Missile Defense
JFCC-ISR	Joint Functional Component Command for Intelligence, Surveillance, and Reconnaissance
JFCH	joint force chaplain
JFE	joint fires element
JFHQ	joint force headquarters
JFHQ–NCR	Joint Force Headquarters–National Capital Region
JFHQ–state	Joint Force Headquarters–State
JFIIT	Joint Fires Integration and Interoperability Team
JFIP	Japanese facilities improvement project
JFLCC	joint force land component commander
JFMC	joint fleet mail center
JFMCC	joint force maritime component commander
JFMO	Joint Frequency Management Office
JFO	joint field office
JFP	joint force package (packaging)
JFRB	Joint Foreign Release Board
JFRG	joint force requirements generator
JFRG II	joint force requirements generator II
JFS	joint force surgeon
JFSOC	joint force special operations component
JFSOCC	joint force special operations component commander
JFTR	joint Federal travel regulations
JFUB	Joint Facilities Utilization Board
JGAT	joint guidance, apportionment, and targeting
JI	joint inspection
JIACG	joint interagency coordination group
JIADS	joint integrated air defense system
JIATF	joint interagency task force (DOD)
JIATF-E	joint interagency task force-East

JIATF-S	joint interagency task force-South
JIATF-W	joint interagency task force-West
JIB	joint information bureau
JIC	joint information center
JICC	joint information coordination center; joint interface control cell
JICO	joint interface control officer
JICPAC	Joint Intelligence Center, Pacific
JICTRANS	Joint Intelligence Center for Transportation
JIDC	joint intelligence and debriefing center; joint interrogation and debriefing center
JIEO	joint interoperability engineering organization
JIEP	joint intelligence estimate for planning
JIES	joint interoperability evaluation system
JIG	joint interrogation group
JILE	joint intelligence liaison element
JIMB	joint information management board
JIMP	joint implementation master plan
JIMPP	joint industrial mobilization planning process
JIMS	joint information management system
JINTACCS	Joint Interoperability of Tactical Command and Control Systems
JIO	joint interrogation operations
JIOC	joint information operations center; joint intelligence operations center
JIOCPAC	Joint Intelligence Operations Center, Pacific
JIOC-SOUTH	Joint Intelligence Operations Center, South
JIOC TRANS	Joint Intelligence Operations Center, Transportation
JIOP	joint interface operational procedures
JIOP-MTF	joint interface operating procedures-message text formats
JIOWC	Joint Information Operations Warfare Command
JIPB	joint intelligence preparation of the battlespace
JIPC	joint imagery production complex
JIPCL	joint integrated prioritized collection list
JIPOE	joint intelligence preparation of the operational environment
JIPTL	joint integrated prioritized target list
JIS	joint information system
JISE	joint intelligence support element
JITC	joint interoperability test command
JITF-CT	Joint Intelligence Task Force for Combating Terrorism
JIVA	Joint Intelligence Virtual Architecture
JKDDC	Joint Knowledge Development and Distribution Capability
JLCC	joint lighterage control center; joint logistics coordination center
JLLP	Joint Lessons Learned Program
JLNCHREP	joint launch report
JLOC	joint logistics operations center
JLOTS	joint logistics over-the-shore
JLRC	joint logistics readiness center
JLSB	joint line of communications security board
JLSE	joint legal support element
JM&S	joint modeling and simulation
JMAARS	joint model after-action review system
JMAG	Joint METOC Advisory Group

JMAO	joint mortuary affairs office; joint mortuary affairs officer
JMAR	joint medical asset repository
JMAS	joint manpower automation system
JMAT	joint mobility assistance team
JMC	joint military command; joint movement center
JMCG	Joint Mobility Control Group; joint movement control group
JMCIS	joint maritime command information system
JMCOMS	joint maritime communications system
JMD	joint manning document
JMeDSAF	joint medical semi-automated forces
JMEM	Joint Munitions Effectiveness Manual
JMEM-SO	Joint Munitions Effectiveness Manual-Special Operations
JMET	joint mission-essential task
JMETL	joint mission-essential task list
JMFU	joint meteorological and oceanographic (METOC) forecast unit
JMIC	Joint Military Intelligence College
JMICS	Joint Worldwide Intelligence Communications System (JWICS) mobile integrated communications system
JMIE	joint maritime information element
JMIP	joint military intelligence program
JMITC	Joint Military Intelligence Training Center
JMLO	joint medical logistics officer
JMMC	Joint Material Management Center
JMMT	joint military mail terminal
JMNA	joint military net assessment
JMO	joint force meteorological and oceanographic officer; Joint maritime operations
JMO(AIR)	joint maritime operations (air)
JMOC	joint medical operations center
JMP	joint manpower program
JMPA	joint military postal activity; joint military satellite communications (MILSATCOM) panel administrator
JMPAB	Joint Materiel Priorities and Allocation Board
JMRC	joint mobile relay center
JMRO	Joint Medical Regulating Office
JMRR	Joint Monthly Readiness Review
JMSEP	joint modeling and simulation executive panel
JMSWG	Joint Multi-Tactical Digital Information Link (Multi-TADIL) Standards Working Group
JMT	joint military training
JMTCA	joint munitions transportation coordinating activity
JMTCSS	Joint Maritime Tactical Communications Switching System
JMUA	Joint Meritorious Unit Award
JMV	joint METOC viewer
JMWG	joint medical working group
JNACC	joint nuclear accident coordinating center
JNCC	joint network operations (NETOPS) control center
JNOCC	Joint Operation Planning and Execution System (JOPES) Network Operation Control Center
JNPE	joint nuclear planning element

JOA	joint operations area
JOAF	joint operations area forecast
JOC	joint operations center; joint oversight committee
JOCC	joint operations command center
JOG	joint operations graphic
JOGS	joint operation graphics system
JOPES	Joint Operation Planning and Execution System
JOPESIR	Joint Operation Planning and Execution System Incident Reporting System
JOPESREP	Joint Operation Planning and Execution System Reporting System
JOPP	joint operation planning process
JOR	joint operational requirement
JORD	joint operational requirements document
JOSG	joint operational steering group
JOT&E	joint operational test and evaluation
JOTS	Joint Operational Tactical System
JP	joint publication
JPAC	joint planning augmentation cell; Joint POW/MIA Accounting Command
JPAG	Joint Planning Advisory Group
JPASE	Joint Public Affairs Support Element
JPATS	joint primary aircraft training system
JPAV	joint personnel asset visibility
JPC	joint planning cell; joint postal cell
JPD	joint planning document
JPEC	joint planning and execution community
JPERSTAT	joint personnel status and casualty report
JPG	joint planning group
JPME	joint professional military education
JPMRC	joint patient movement requirements center
JPMT	joint patient movement team
JPN	joint planning network
JPO	Joint Petroleum Office; Joint Program Office
JPOC	joint planning orientation course
JPOI	joint program of instruction
JPOM	joint preparation and onward movement
JPO-STC	Joint Program Office for Special Technology Countermeasures
JPOTF	joint psychological operations task force
JPOTG	joint psychological operations task group
JPRA	Joint Personnel Recovery Agency
JPRC	joint personnel receiving center; joint personnel reception center; joint personnel recovery center
JPRSP	joint personnel recovery support product
JPS	joint processing system
JPTTA	joint personnel training and tracking activity
JQR	joint qualification requirements
JQRR	joint quarterly readiness review
JRADS	Joint Resource Assessment Data System
JRB	Joint Requirements Oversight Council (JROC) Review Board
JRC	joint reception center; joint reconnaissance center

JRCC	joint reception coordination center
JRERP	Joint Radiological Emergency Response Plan
JRFL	joint restricted frequency list
JRG	joint review group
JRIC	joint reserve intelligence center
JRMB	Joint Requirements and Management Board
JROC	Joint Requirements Oversight Council
JRS	joint reporting structure
JRSC	jam-resistant secure communications; joint rescue sub-center
JRSOI	joint reception, staging, onward movement, and integration
JRTC	joint readiness training center
JRX	joint readiness exercise
JS	the joint staff
JSA	joint security area
JSAC	joint strike analysis cell; joint strike analysis center
JSAM	joint security assistance memorandum; Joint Service Achievement Medal; joint standoff surface attack missile
JSAN	Joint Staff automation for the nineties
JSAP	Joint Staff action process
JSAS	joint strike analysis system
JSC	joint security coordinator; Joint Spectrum Center
JSCAT	joint staff crisis action team
JSCC	joint security coordination center; joint Services coordination committee
JSCM	joint Service commendation medal
JSCP	Joint Strategic Capabilities Plan
JSDS	Joint Staff doctrine sponsor
J-SEAD	joint suppression of enemy air defenses
JSEC	joint strategic exploitation center
JSIDS	joint Services imagery digitizing system
JSIR	joint spectrum interference resolution
JSISC	Joint Staff Information Service Center
JSIT	Joint Operation Planning and Execution System (JOPES) information trace
JSIVA	Joint Staff Integrated Vulnerability Assessment
JSM	Joint Staff Manual
JSME	joint spectrum management element
JSMS	joint spectrum management system
JSO	joint security operations; joint specialty officer or joint specialist
JSOA	joint special operations area
JSOAC	joint special operations air component; joint special operations aviation component
JSOACC	joint special operations air component commander
JSOC	joint special operations command
JSOFI	Joint Special Operations Forces Institute
JSOTF	joint special operations task force
JSOU	Joint Special Operations University
JSOW	joint stand-off weapon
JSPA	joint satellite communications (SATCOM) panel administrator
JSPD	joint strategic planning document

JSPDSA	joint strategic planning document supporting analyses
JSPOC	Joint Space Operations Center
JSPS	Joint Strategic Planning System
JSR	joint strategy review
JSRC	joint subregional command (NATO)
JSS	joint surveillance system
JSSA	joint Services survival, evasion, resistance, and escape (SERE) agency
JSSIS	joint staff support information system
JSST	joint space support team
JSTAR	joint system threat assessment report
JSTARS	Joint Surveillance Target Attack Radar System
JSTE	joint system training exercise
JT&E	joint test and evaluation
JTA	joint table of allowances; joint technical architecture
JTAC	joint technical augmentation cell; joint terminal attack controller; Joint Terrorism Analysis Center
JTADS	Joint Tactical Air Defense System (Army); Joint Tactical Display System
JTAGS	joint tactical ground station (Army); joint tactical ground station (Army and Navy); joint tactical ground system
JTAO	joint tactical air operations
JTAR	joint tactical air strike request
JTASC	joint training analysis and simulation center
JTASG	Joint Targeting Automation Steering Group
JTAV	joint total asset visibility
JTAV-IT	joint total asset visibility-in theater
JTB	Joint Transportation Board
JTC	joint technical committee; Joint Training Confederation
JTCB	joint targeting coordination board
JTCC	joint transportation coordination cell; joint transportation corporate information management center
JTCG/ME	Joint Technical Coordinating Group for Munitions Effectiveness
JTD	joint table of distribution; joint theater distribution
JTDC	joint track data coordinator
JTF	joint task force
JTF-6	joint task force-6
JTF-AK	Joint Task Force - Alaska
JTF-B	joint task force-Bravo
JTFCEM	joint task force contingency engineering management
JTF-CM	joint task force - consequence management
JTF-CS	Joint Task Force-Civil Support
JTF-GNO	Joint Task Force-Global Network Operations
JTF-GTMO	Joint Task Force-Guantanamo
JTF-HD	Joint Task Force-Homeland Defense
JTF HQ	joint task force headquarters
JTF-MAO	joint task force - mortuary affairs office
JTF-N	Joint Task Force-North
JTFP	Joint Tactical Fusion Program
JTFS	joint task force surgeon
JTF-State	Joint Task Force-State
JTIC	joint transportation intelligence center

JTIDS	Joint Tactical Information Distribution System
JTL	joint target list
JTLM	Joint Theater Logistics Management
JTLS	joint theater-level simulation
JTM	joint training manual
JTMD	joint table of mobilization and distribution; Joint Terminology Master Database
JTMP	joint training master plan
JTMS	joint theater movement staff; joint training master schedule
JTP	joint test publication; joint training plan
JTR	joint travel regulations
JTRB	joint telecommunication resources board
JTS	Joint Targeting School
JTSG	joint targeting steering group
JTSSCCB	Joint Tactical Switched Systems Configuration Control Board
JTSST	joint training system support team
JTT	joint targeting toolbox; joint training team
JTTF	joint terrorism task force
JUH-MTF	Joint User Handbook-Message Text Formats
JUIC	joint unit identification code
JULL	Joint Universal Lessons Learned (report)
JULLS	Joint Universal Lessons Learned System
JUO	joint urban operation
JUSMAG	Joint United States Military Advisory Group
JUWTF	joint unconventional warfare task force
JV	Joint Vision
JV 2020	Joint Vision 2020
JVB	Joint Visitors Bureau
JVIDS	Joint Visual Integrated Display System
JVSEAS	Joint Virtual Security Environment Assessment System
JWAC	Joint Warfare Analysis Center
JWARS	Joint Warfare Analysis and Requirements System
JWC	Joint Warfare Center
JWCA	joint warfighting capabilities assessment
JWFC	Joint Warfighting Center
JWG	joint working group
JWICS	Joint Worldwide Intelligence Communications System
JWID	joint warrior interoperability demonstration

Ka	Kurtz-above band
KAL	key assets list
KAPP	Key Assets Protection Program
kb	kilobit
kbps	kilobits per second
KC-135	Stratotanker
KDE	key doctrine element
KEK	key encryption key
KG	key generator
kg	kilogram
kHz	kilohertz

KIA	killed in action
K-Kill	catastrophic kill
km	kilometer
KMC	knowledge management center
KNP	Korean National Police
KP	key pulse
kph	kilometers per hour
KQ	tactical location identifiers
kt	kiloton(s); knot (nautical miles per hour)
Ku	Kurtz-under band
kVA	kilo Volt-Amps
KVG	key variable generator
kW	kilowatt
KWOC	keyword-out-of-context
L	length
l	search subarea length
LA	lead agent; legal adviser; line amplifier; loop key generator (LKG) adapter
LAADS	low altitude air defense system
LAAM	light anti-aircraft missile
LABS	laser airborne bathymetry system
LACH	lightweight amphibious container handler
LACV	lighter, air cushioned vehicle
LAD	latest arrival date
LAMPS	Light Airborne Multipurpose System (helicopter)
LAN	local area network
LANDCENT	Allied Land Forces Central Europe (NATO)
LANDSAT	land satellite
LANDSOUTH	Allied Land Forces Southern Europe (NATO)
LANTIRN	low-altitude navigation and targeting infrared for night
LAO	limited attack option
LAPES	low-altitude parachute extraction system
LARC	lighter, amphibious resupply, cargo
LARC-V	lighter, amphibious resupply, cargo, 5 ton
LARS	lightweight airborne recovery system
LASH	lighter aboard ship
LASINT	laser intelligence
LAT	latitude
LAV	light armored vehicle
lb	pound
LBR	Laser Beam Rider
LC	lake current; legal counsel
LCAC	landing craft, air cushion
LCAP	low combat air patrol
LCB	line of constant bearing
LCC	amphibious command ship; land component commander; launch control center; lighterage control center; link communications circuit; logistics component command
LCCS	landing craft control ship

LCE	logistics capability estimator; logistics combat element
LCES	line conditioning equipment scanner
LCM	landing craft, mechanized; letter-class mail; life-cycle management
LCO	lighterage control officer
LCP	lighterage control point
LCPL	landing craft personnel (large)
LCU	landing craft, utility; launch correlation unit
LCVP	landing craft, vehicle, personnel
LD	line of departure
LDA	limited depository account
LDF	lightweight digital facsimile
LDI	line driver interface
LDO	laser designator operator
LDR	leader; low data rate
LE	law enforcement; low-order explosives
LEA	law enforcement agency
LEAP	Light ExoAtmospheric Projectile
LEASAT	leased satellite
LEAU	Law Enforcement Assistance Unit (FAA)
LECIC	Law Enforcement and Counterintelligence Center (DOD)
LED	law enforcement desk; light emitting diode
LEDET	law enforcement detachment
LEGAT	legal attaché
LEO	law enforcement operations; low Earth orbit
LEP	laser eye protection; linear error probable
LERSM	lower echelon reporting and surveillance module
LERTCON	alert condition
LES	law enforcement sensitive; leave and earnings statement; Lincoln Laboratories Experimental Satellite
LESO	Law Enforcement Support Office
LET	light equipment transport
LF	landing force; low frequency
LFA	lead federal agency
LFORM	landing force operational reserve material
LFSP	landing force support party
LfV	*Landesamt für Verfassungsschutz* (regional authority for constitutional protection)
LG	deputy chief of staff for logistics
LGB	laser-guided bomb
LGM	laser-guided missile; loop group multiplexer
LGM-30	Minuteman
LGW	laser-guided weapon
LHA	amphibious assault ship (general purpose); amphibious assault ship (multi-purpose)
LHD	amphibious assault ship (dock)
L-hour	specific hour on C-day at which a deployment operation commences or is to commence
LHT	line-haul tractor
LIDAR	light detection and ranging
LIF	light interference filter

LIMDIS	limited distribution
LIMFAC	limiting factor
LIPS	Logistics Information Processing System
LIS	logistics information system
LIWA	land information warfare activity
LKA	attack cargo ship
LKG	loop key generator
LKP	last known position
LL	lessons learned
LLLGB	low-level laser-guided bomb
LLLTV	low-light level television
LLSO	low-level source operation
LLTR	low-level transit route
LM	loop modem
LMARS	Logistics Metrics Analysis Reporting System
LMAV	laser MAVERICK
LMF	language media format
LMSR	large, medium speed roll-on/roll-off
LMW	lead mobility wing
LN	lead nation
LNA	low voice amplifier
LNO	liaison officer
LO	low observable
LOA	Lead Operational Authority; letter of assist; letter of authorization; letter of offer and acceptance; lodgment operational area; logistics over-the-shore (LOTS) operation area
LOAC	law of armed conflict
LOAL	lock-on after launch
LOBL	lock-on before launch
LOC	line of communications; logistics operations center
LOC ACC	location accuracy
LOCAP	low combat air patrol
LOCE	Linked Operational Intelligence Centers Europe; Linked Operations-Intelligence Centers Europe
LOE	letter of evaluation
LOG	logistics
LOGAIR	logistics aircraft
LOGAIS	logistics automated information system
LOGCAP	logistics civil augmentation program
LOGCAT	logistics capability assessment tool
LOGDET	logistics detail
LOGEX	logistics exercise
LOGFAC	Logistics Feasibility Assessment Capability
LOGFOR	logistics force packaging system
LOGMARS	logistics applications of automated marking and reading symbols
LOGMOD	logistics module
LOGPLAN	logistics planning system
LOGSAFE	logistic sustainment analysis and feasibility estimator
LOI	letter of instruction; loss of input
LO/LO	lift-on/lift-off

LOMEZ	low-altitude missile engagement zone
LONG	longitude
LOO	line of operations
LOP	line of position
LORAN	long-range aid to navigation
LO/RO	lift-on/roll-off
LOROP	long range oblique photography
LOS	line of sight
LOTS	logistics over-the-shore
LOX	liquid oxygen
LP	listening post
LPD	amphibious transport dock; low probability of detection
LPH	amphibious assault ship, landing platform helicopter
LPI	low probability of intercept
LPSB	logistics procurement support board
LPU	line printer unit
LPV	laser-protective visor
LRC	logistics readiness center
LRD	laser range finder-detector
LRF	laser range finder
LRF/D	laser range finder/detector
LRG	long-range aircraft
LRM	low rate multiplexer
LRP	load and roll pallet
LRRP	long range reconnaissance patrol
LRS	launch and recovery site
LRST	long-range surveillance team
LRSU	long-range surveillance unit
LSA	logistic support analysis
LSB	landing support battalion; lower sideband
LSCDM	low speed cable driver modem
LSD	landing ship dock; least significant digit
LSE	landing signal enlisted; logistic support element
LSO	landing safety officer; landing signal officer
LSPR	low speed pulse restorer
LSS	local sensor subsystem
LST	landing ship, tank; laser spot tracker; tank landing ship
LSV	logistics support vessel
LT	large tug;local terminal; long ton
L/T	long ton
LTD	laser target designator
LTD/R	laser target designator/ranger
LTF	logistics task force
LTG	local timing generator
LTON	long ton
LTS	low-altitude navigation and targeting infrared for night (LANTIRN) targeting system
LTT	loss to theater
LTU	line termination unit
LUA	launch under attack

LUT	local user terminal
LVS	Logistics Vehicle System (USMC)
LW	leeway
LWR	Lutheran World Relief
LZ	landing zone
LZCO	landing zone control officer
M&S	modeling and simulation
M88A1	recovery vehicle
MA	master; medical attendant; mortuary affairs
mA	milliampere(s)
MAAG	military assistance advisory group
MAAP	master air attack plan
MAC	Mortuary Affairs Center
MACA	military assistance to civil authorities
MACB	multinational acquisition and contracting board
MACCS	Marine air command and control system
MACDIS	military assistance for civil disturbances
MACG	Marine air control group
MACOM	major command (Army)
MACP	mortuary affairs collection point
MACSAT	multiple access commercial satellite
MAD	*Militärischer Abschirmdienst* (military protection service); military air distress
MADCP	mortuary affairs decontamination collection point
MAEB	mean area of effectiveness for blast
MAEF	mean area of effectiveness for fragments
MAF	mobility air forces
MAFC	Marine air-ground task force (MAGTF) all-source fusion center
MAG	Marine aircraft group
MAGTF	Marine air-ground task force
MAGTF ACE	Marine air-ground task force aviation combat element
MAJCOM	major command (USAF)
MANFOR	manpower force packaging system
MANPADS	man-portable air defense system
MANPER	manpower and personnel module
MAOC-N	Maritime Analysis and Operations Center-Narcotics
MAP	Military Assistance Program; missed approach point; missed approach procedure
MAR	METOC assistance request
MARAD	Maritime Administration
MARCORMATCOM	Marine Corps Materiel Command
MARDIV	Marine division
MARFOR	Marine Corps forces
MARFOREUR	Marine Corps Forces, Europe
MARFORLANT	Marine Corps Forces, Atlantic
MARFORNORTH	Marine Corps Forces, North
MARFORPAC	Marine Corps Forces, Pacific
MARFORSOUTH	Marine Corps Forces, South
MARINCEN	Maritime Intelligence Center

MARLO	Marine liaison officer
MAROP	marine operators
MARPOL	International Convention for the Prevention of Pollution from Ships
MARS	Military Affiliate Radio System
MARSA	military assumes responsibility for separation of aircraft
MARSOC	Marine Corps special operations command
MARSOF	Marine Corps special operations forces
MART	mobile Automatic Digital Network (AUTODIN) remote terminal
MASCAL	mass casualty
MASF	mobile aeromedical staging facility
MASH	mobile Army surgical hospital
MASINT	measurement and signature intelligence
MASLO	measurement and signature intelligence (MASINT) liaison officer
MAST	military assistance to safety and traffic; mobile ashore support terminal
MAT	medical analysis tool
MATCALS	Marine air traffic control and landing system
M/ATMP	Missiles/Air Target Materials Program
MAW	Marine aircraft wing
MAX	maximum
MAXORD	maximum ordinate
MB	medium boat; megabyte
MBA	main battle area
MBBLs	thousands of barrels
MBCDM	medical biological chemical defense materiel
MBI	major budget issue
Mbps	megabytes per second
Mbs	megabits per second
MC	Military Committee (NATO); military community; missioncapable
MC-130	Combat Talon (I and II)
MCA	mail control activity; maximum calling area; military civic action; mission concept approval; movement control agency
MCAP	maximum calling area precedence
MCAS	Marine Corps air station
MCB	movement control battalion
MCBAT	medical chemical biological advisory team
MCC	Marine component commander; maritime component commander; master control center; military cooperation committee; military coordinating committee; mission control center; mobility control center; movement control center
MCCC	mobile consolidated command center
MCCDC	Marine Corps Combat Development Command
MCCISWG	military command, control, and information systems working group
MCD	medical crew director
MCDA	military and civil defense assets (UN)
MCDP	Marine Corps doctrinal publication
MCDS	modular cargo delivery system
MCEB	Military Communications-Electronics Board
MCEWG	Military Communications-Electronics Working Group
MC/FI	mass casualty/fatality incident
MCIA	Marine Corps Intelligence Activity

MCIO	military criminal investigation organization
MCIP	Marine Corps information publication; military command inspection program
MCM	Manual for Courts-Martial; military classification manual; mine countermeasures
MCMC	mine countermeasures commander
MCMG	Military Committee Meteorological Group (NATO)
MCMO	medical civil-military operations
MCMOPS	mine countermeasures operations
M/CM/S	mobility, countermobility, and/or survivability
MCMREP	mine countermeasure report
MCO	Mapping Customer Operations; Marine Corps order
MCOO	modified combined obstacle overlay
MCRP	Marine Corps reference publication
MCS	maneuver control system; Military Capabilities Study; Mine countermeasures ship; modular causeway system
MCSF	mobile cryptologic support facility
MCSFB	Marine Corps security force battalion
MCT	movement control team
MCTC	Midwest Counterdrug Training Center
MCTFT	Multijurisdictional Counterdrug Task Force Training
MCU	maintenance communications unit
MCW	modulated carrier wave
MCWP	Marine Corps warfighting publication
MCX	Marine Corps Exchange
MDA	Magen David Adom (Israeli equivalent of the Red Cross); maritime domain awareness
M-DARC	military direct access radar channel
M-day	mobilization day; unnamed day on which mobilization of forces begins
MDCI	multidiscipline counterintelligence
MDDOC	MAGTF deployment and distribution operations center
MDF	Main Defense Forces (NATO); main distribution frame
MDITDS	migration defense intelligence threat data system; Modernized Defense Intelligence Threat Data System
MDMA	methylenedioxymethamphetamine
MDR	medium data rate
MDRO	mission disaster response officer
MDS	Message Dissemination Subsystem; mission design series
MDSS II	Marine air-ground task force (MAGTF) Deployment Support System II
MDSU	mobile diving and salvage unit
MDW	Military District of Washington
MDZ	maritime defense zone
MEA	munitions effect assessment; munitions effectiveness assessment
MEB	Marine expeditionary brigade
MEBU	mission essential backup
MEC	medium endurance cutter
ME/C	medical examiner and/or coroner
MED	manipulative electronic deception
MEDAL	Mine Warfare Environmental Decision Aids Library
MEDCAP	medical civic action program

MEDCC	medical coordination cell
MEDCOM	US Army Medical Command
MEDEVAC	medical evacuation
MEDINT	medical intelligence
MEDLOG	medical logistics (USAF AIS)
MEDLOGCO	medical logistics company
MEDLOG JR	medical logistics, junior (USAF)
MEDMOB	Medical Mobilization Planning and Execution System
MEDNEO	medical noncombatant evacuation operation
MEDREG	medical regulating
MEDREGREP	medical regulating report
MEDRETE	medical readiness training exercise
MEDS	meteorological data system
MEDSOM	medical supply, optical, and maintenance unit
MEDSTAT	medical status
MEF	Marine expeditionary force
MEF(FWD)	Marine expeditionary force (forward)
MEFPAKA	manpower and equipment force packaging
MEL	maintenance expenditure limit; minimum equipment list
MEO	medium Earth orbit; military equal opportunity
MEP	mobile electric power
MEPCOM	military entrance processing command
MEPES	Medical Planning and Execution System
MEPRS	Military Entrance Processing and Reporting System
MERCO	merchant ship reporting and control
MERINT	merchant intelligence
MERSHIPS	merchant ships
MES	medical equipment set
MESAR	minimum-essential security assistance requirements
MET	medium equipment transporter; mobile environmental team
METAR	meteorological airfield report; meteorological aviation report
METARS	routine aviation weather report (roughly translated from French); international standard code format for hourly surface weather observations)
METCON	control of meteorological information (roughly translated from French); meteorological control (Navy)
METL	mission-essential task list
METMF	meteorological mobile facility
METMR(R)	meteorological mobile facility (replacement)
METOC	meteorological and oceanographic
METSAT	meteorological satellite
METT-T	mission, enemy, terrain and weather, troops and support available—time available
METT-TC	mission, enemy, terrain and weather, troops and support available-time available and civil considerations (Army)
MEU	Marine expeditionary unit
MEU(SOC)	Marine expeditionary unit (special operations capable)
MEVA	mission essential vulnerable area
MEWSG	Multi-Service Electronic Warfare Support Group (NATO)
MEZ	missile engagement zone

MF	medium frequency; mobile facility; multi-frequency
MFC	Meteorological and Oceanographic (METOC) Forecast Center; multinational force commander
MFDS	Modular Fuel Delivery System
MFE	manpower force element
MFFIMS	mass fatality field information management system
MFO	multinational force and observers
MFP	major force program
MFPC	maritime future plans center
MFPF	minefield planning folder
MFS	multifunction switch
MGB	medium girder bridge
MGM	master group multiplexer
MGRS	military grid reference system
MGS	mobile ground system
MGT	management
MHC	management headquarters ceiling
MHE	materials handling equipment
MHU	modular heat unit
MHW	mean high water
MHz	megahertz
MI	military intelligence; movement instructions
MIA	missing in action
MIAC	maritime intelligence and analysis center
MIB	Military Intelligence Board
MIC	Multinational Interoperability Council
MICAP	mission capable/mission capability
MICON	mission concept
MICRO-MICS	micro-medical inventory control system
MICRO-SNAP	micro-shipboard non-tactical automated data processing system
MIDB	modernized integrated database; modernized intelligence database
MIDDS-T	Meteorological and Oceanographic (METOC) Integrated Data Display System-Tactical
MIF	maritime interception force
MIIDS/IDB	Military Intelligence Integrated Data System/Integrated Database
MIJI	meaconing, interference, jamming, and intrusion
MILALOC	military air line of communications
MILCON	military construction
MILDEC	military deception
MILDEP	Military Department
MILGP	military group (assigned to American Embassy in host nation)
MILOB	military observer
MILOC	military oceanography group (NATO)
MILPERS	military personnel
MILSATCOM	military satellite communications
MILSPEC	military performance specification
MILSTAMP	military standard transportation and movement procedures
MILSTAR	military strategic and tactical relay system
MIL-STD	military standard
MILSTRAP	military standard transaction reporting and accounting procedure

MILSTRIP	military standard requisitioning and issue procedure
MILTECH	military technician
MILU	multinational integrated logistic support unit
MILVAN	military van (container)
MIM	maintenance instruction manual
MIMP	Mobilization Information Management Plan
MINEOPS	joint minelaying operations
MIO	maritime interception operations
MIO-9	information operations threat analysis division (DIA)
MIP	Military Intelligence Program
MIPE	mobile intelligence processing element
MIPOE	medical intelligence preparation of the operational environment
MIPR	military interdepartmental purchase request
MIS	maritime intelligence summary
MISCAP	mission capability
MISREP	mission report
MISS	missing
MIST	military information support team
MITASK	mission tasking
MITO	minimum interval takeoff
MITT	mobile integrated tactical terminal
MIUW	mobile inshore undersea warfare
MIUWU	mobile inshore undersea warfare unit
MIW	mine warfare
MJCS	Joint Chiefs of Staff memorandum
MJLC	multinational joint logistic center
M-Kill	mobility kill
MLA	mission load allowance
MLAYREP	mine laying report
MLC	Marine Logistics Command
MLE	maritime law enforcement
MLEA	Maritime Law Enforcement Academy
MLI	munitions list item
MLMC	medical logistics management center
MLO	military liaison office
MLP	message load plan
MLPP	multilevel precedence and preemption
MLPS	Medical Logistics Proponent Subcommittee
MLRS	Multiple Launch Rocket System
MLS	microwave landing system; multilevel security
MLW	mean low water
MMAC	military mine action center
MMC	materiel management center
MMG	DOD Master Mobilization Guide
MMI	man/machine interface
MMLS	mobile microwave landing system
MMS	mast-mounted sight
MMT	military mail terminal
MNCC	multinational coordination center
MNF	multinational force

MNFACC	multinational force air component commander
MNFC	multinational force commander
MNFLCC	multinational force land component commander
MNFMCC	multinational force maritime component commander
MNFSOCC	multinational force special operations component commander
MNL	multinational logistics
MNLC	multinational logistic center
MNP	master navigation plan
MNS	mine neutralization system (USN); mission needs statement
MNTF	multinational task force
MO	medical officer; month
MOA	memorandum of agreement; military operating area
MOADS	maneuver-oriented ammunition distribution system
MOB	main operating base; main operations base; mobilization
MOBCON	mobilization control
MOBREP	military manpower mobilization and accession status report
MOC	maritime operations center
MOCC	measurement and signature intelligence (MASINT) operations coordination center; mobile operations control center
MOD	Minister (Ministry) of Defense
MODEM	modulator/demodulator
MODLOC	miscellaneous operational details, local operations
MOD T-AGOS	modified tactical auxiliary general ocean surveillance
MOE	measure of effectiveness
MOG	maximum (aircraft) on ground; movement on ground (aircraft); multinational observer group
MOGAS	motor gasoline
MOLE	multichannel operational line evaluator
MOM	military ordinary mail
MOMAT	mobility matting
MOMSS	mode and message selection system
MOP	measure of performance; memorandum of policy
MOPP	mission-oriented protective posture
MOR	memorandum of record
MOS	military occupational specialty
MOTR	maritime operational threat response
MOU	memorandum of understanding
MOUT	military operations in urban terrain; ilitary operations on urbanized terrain
MOVREP	movement report
MOW	maintenance orderwire
MP	military police (Army and Marine); multinational publication
MPA	maritime patrol aircraft; mission and payload assessment; mission planning agent
MPAT	military patient administration team; Multinational Planning Augmentation Team
MPC	mid-planning conference; military personnel center
MPE/S	maritime pre-positioning equipment and supplies
MPF	maritime pre-positioning force
MPG	maritime planning group; mensurated point graphic

mph	miles per hour
MPLAN	Marine Corps Mobilization Management Plan
MPM	medical planning module
MPO	military post office
MPR	maritime patrol and reconnaissance
MPRS	multi-point refueling system
MPS	maritime pre-positioning ship; message processor shelter; Military Postal Service
MPSA	Military Postal Service Agency
MPSRON	maritime pre-positioning ships squadron
MR	milliradian; mobile reserve
MRAALS	Marine remote area approach and landing system
MRAP	Mine Resistant Ambush Protected Vehicle ission Readiness Assessment Tool
MRAT	medical radiobiology advisory team
MRCI	maximum rescue coverage intercept
MRE	meal, ready to eat
MRG	movement requirements generator
MRI	magnetic resonance imaging
MRMC	US Army Medical Research and Materiel Command
MRO	mass rescue operation; medical regulating office; medical regulating officer
MROC	multicommand required operational capability
MRR	minimum-risk route
MRRR	mobility requirement resource roster
MRS	measurement and signature intelligence (MASINT) requirements system; meteorological radar subsystem; movement report system
MRSA	Materiel Readiness Support Agency
MRT	maintenance recovery team
MRU	mountain rescue unit
MS	message switch
ms	millisecond
MSC	major subordinate command; maritime support center; Military Sealift Command; military staff committee; mission support confirmation
MSC	military support to civil authorities; military support to civilian authorities
MSCD	military support to civil defense
MSCLEA	military support to civilian law enforcement agencies
MSCO	Military Sealift Command Office
MSD	marginal support date; mobile security division
MS-DOS	Microsoft disk operating system
MSDS	mission specific data set
MSE	mission support element; mobile subscriber equipment
MSECR	HIS 6000 security module
MSEL	master scenario events list
MSF	*Medicins Sans Frontieres* (Doctors Without Borders); mission support force; mobile security force; multiplex signal format
MSG	Marine Security Guard; message
MSGID	message identification

MSHARPP	mission, symbolism, history, accessibility, recognizability, population, and proximity
MSI	modified surface index; multi-spectral imagery
MSIC	Missile and Space Intelligence Center
MSIS	Marine safety information system
MSK	mission support kit
MSL	master station log; mean sea level
MSNAP	merchant ship naval augmentation program
MSO	map support office; marine safety office(r); military satellite communications (MILSATCOM) systems organization; military source operation; military support operations; mobilization staff officer
MSOC	Marine special operations company
MSP	mission support plan; mobile sensor platform
MSPES	mobilization stationing, planning, and execution system
MSPF	maritime special purpose force
MSPS	mobilization stationing and planning system
MSR	main supply route; maritime support request; mission support request
MSRR	modeling and simulation resource repository
MSRV	message switch rekeying variable
MSS	medical surveillance system; meteorological satellite subsystem
MSSG	Marine expeditionary unit (MEU) service support group
MST	Marine expeditionary force (MEF) weather support team; mission support team
M/T	measurement ton
MT	measurement ton; military technician ministry team
MTA	military training agreement
MTAC	Multiple Threat Alert Center
MTBF	mean time between failures
MT Bn	motor transport battalion
MT/D	measurement tons per day
MTF	medical treatment facility; message text format
MTG	master timing generator
MTI	moving target indicator
MTIC	Military Targeting Intelligence Committee
MTL	mission tasking letter
MTMS	maritime tactical message system
MTN	multi-tactical data link network
MTO	message to observer mission type order
MTOE	modified table of organization and equipment
MTON	measurement ton
MTP	maritime task plan; mission tasking packet
MTS	Movement Tracking System
MTS/SOF-IRIS	multifunction system
MTT	magnetic tape transport; mobile training team
MTW	major theater war
MTX	message text format
MU	marry up
MUL	master urgency list (DOD)
MULE	modular universal laser equipment
MUREP	munitions report

MUSARC	major United States Army reserve commands
MUSE	mobile utilities support equipment
MUST	medical unit, self-contained, transportable
MUX	multiplex
MV	merchant vessel; motor vessel
mV	millivolt
MWBP	missile warning bypass
MWC	Missile Warning Center (NORAD)
MWD	military working dog
MWDT	military working dog team
MWF	medical working file
MWG	mobilization working group
MWOD	multiple word-of-day
MWR	missile warning receiver; morale, welfare, and recreation
MWSG	Marine wing support group
MWSS	Marine wing support squadron
N	number of required track spacings number of search and rescue units (SRUs)
N-1	Navy component manpower or personnel staff officer
N-2	Director of Naval Intelligence; Navy component intelligence staff officer
N-3	Navy component operations staff officer
N-4	Navy component logistics staff officer
N-5	Navy component plans staff officer
N-6	Navy component communications staff officer
NAAG	North Atlantic Treaty Organization (NATO) Army Armaments Group
NAC	North American Aerospace Defense Command (NORAD) Air Center; North Atlantic Council (NATO)
NACE	National Military Command System (NMCS) Automated Control Executive
NACISA	North Atlantic Treaty Organization (NATO) Communications and Information Systems Agency
NACISC	North Atlantic Treaty Organization (NATO) Communications and Information Systems Committee
NACSEM	National Communications Security/Emanations Security (COMSEC/EMSEC) Information Memorandum
NACSI	national communications security (COMSEC) instruction
NACSIM	national communications security (COMSEC) information memorandum
NADEFCOL	North Atlantic Treaty Organization (NATO) Defense College
NADEP	naval aircraft depot
NAEC-ENG	Naval Air Engineering Center - Engineering
NAF	naval air facility; nonappropriated funds; numbered air force
NAFAG	North Atlantic Treaty Organization (NATO) Air Force Armaments Group
NAI	named area of interest
NAIC	National Air Intelligence Center
NAK	negative acknowledgement
NALC	naval ammunition logistic code

NALE	naval and amphibious liaison element
NALSS	naval advanced logistic support site
NAMP	North Atlantic Treaty Organization (NATO) Annual Manpower Plan
NAMS	National Air Mobility System
NAMTO	Navy material transportation office
NAOC	national airborne operations center (E-4B aircraft)
NAPCAP	North Atlantic Treaty Organization (NATO) Allied Pre-Committed Civil Aircraft Program
NAPMA	North Atlantic Treaty Organization (NATO) Airborne Early Warning and Control Program Management Agency
NAPMIS	Navy Preventive Medicine Information System
NAR	nonconventional assisted recovery; notice of ammunition reclassification
NARAC	national atmospheric release advisory capability
NARC	non-automatic relay center
NAS	naval air station
NASA	National Aeronautics and Space Administration
NASAR	National Association for Search and Rescue
NAS computer	national airspace system computer
NASIC	National Air and Space Intelligence Center
NAT	nonair-transportable (cargo)
NATO	North Atlantic Treaty Organization
NATOPS	Naval Air Training and Operating Procedures Standardization
NAU	Narcotics Assistance Unit
NAVAID	navigation aid
NAVAIDS	navigational aids
NAVAIR	naval air; Naval Air Systems Command
NAVAIRSYSCOM	Naval Air Systems Command (Also called NAVAIR)
NAVATAC	Navy Antiterrorism Analysis Center; Navy Antiterrorist Alert Center
NAVCHAPDET	naval cargo handling and port group detachment
NAVCHAPGRU	Navy cargo handling and port group
NAVCOMSTA	naval communications station
NAVEODTECHDIV	Naval Explosives Ordnance Disposal Technology Division
NAVEURMETOCCEN	Naval Europe Meteorology and Oceanography Center
NAVFAC	Naval Facilities Engineering Command
NAVFACENGCOM	Naval Facilities Engineering Command
NAVFAC-X	Naval Facilities Engineering Command-expeditionary
NAVFAX	Navy facsimile
NAVFOR	Navy forces
NAVICECEN	Naval Ice Center
NAVLANTMETOCCEN	Naval Atlantic Meteorology and Oceanography Center
NAVMAG	naval magazine
NAVMED	Navy Medical; Navy medicine
NAVMEDCOMINST	Navy medical command instruction
NAVMEDLOGCOM	Navy Medical Logistical Command
NAVMEDP	Navy medical pamphlet
NAVMETOCCOM	Naval Meteorology and Oceanography Command
NAVMTO	naval military transportation office
NAVOCEANO	Naval Oceanographic Office
NAVORD	naval ordnance

NAVORDSTA	naval ordnance station
NAVPACMETOCCEN	Naval Pacific Meteorology and Oceanography Center
NAVSAFECEN	naval safety center
NAVSAT	navigation satellite
NAVSEA	Naval Sea Systems Command
NAVSEAINST	naval sea instruction
NAVSEALOGCEN	naval sea logistics center
NAVSEASYSCOM	Naval Sea Systems Command
NAVSO	United States Navy Forces, Southern Command
NAVSOC	naval special operations command; naval special operations component; naval special warfare special operations component; Navy special operations component
NAVSOF	naval special operations forces
NAVSPACECOM	Naval Space Command
NAVSPECWARCOM	Naval Special Warfare Command
NAVSPOC	Naval Space Operations Center
NAVSUP	Navy Support Instruction
NAWCAD	Naval Air Warfare Center, Aircraft Division
NB	narrowband
NBC	nuclear, biological, and chemical
NBCCS	nuclear, biological, and chemical (NBC) contamination survivability
NBDP	narrow band direct printing
NBG	naval beach group
NBI	nonbattle injury
NBS	National Bureau of Standards
NBST	narrowband secure terminal
NBVC	Naval Base Ventura County
NC3A	nuclear command, control, and communications (C3) assessment
NCAA	North Atlantic Treaty Organization (NATO) Civil Airlift Agency
NCAGS	naval cooperation and guidance for shipping
NCAPS	naval coordination and protection of shipping
NCB	national central bureau; naval construction brigade
NCC	National Coordinating Center; naval component commander; Navy component command; Navy component commander; network control center; North American Aerospace Defense Command (NORAD) Command Center
NCCS	Nuclear Command and Control System
NCD	net control device
NCDC	National Climatic Data Center
NCESGR	National Committee of Employer Support for the Guard and Reserve
NCF	naval construction force
NCFSU	naval construction force support unit
NCHB	Navy cargo handling battalion
NCHF	Navy cargo handling force
NCIC	National Crime Information Center
NCI&KA	national critical infrastructure and key assets
NCIS	Naval Criminal Investigative Service
NCISRA	Naval Criminal Investigative Service resident agent
NCISRO	Naval Criminal Investigative Service regional office
NCISRU	Naval Criminal Investigative Service resident unit

NCIX	National Counterintelligence Executive
NCMP	Navy Capabilities and Mobilization Plan
NCO	noncombat operations; noncommissioned officer
NCOB	National Counterintelligence Operations Board
NCOIC	noncommissioned officer in charge
NCOS	naval control of shipping
NCP	National Oil and Hazardous Substances Pollution Contingency Plan
NCR	National Capital Region (US) national cryptologic representative; National Security Agency/Central Security Service representative; naval construction regiment
NCRCC	National Capital Region Coordination Center
NCRCG	National Cyber Response Coordination Group
NCRDEF	national cryptologic representative defense
NCR-IADS	National Capital Region - Integrated Air Defense System
NCS	National Clandestine Service; National Communications System; naval control of shipping; net control station
NCSC	National Computer Security Center
NCSE	national intelligence support team (NIST) communications support element
NCT	network control terminal
NCTAMS	naval computer and telecommunications area master station
NCTC	National Counterterrorism Center; North East Counterdrug Training Center
NCTS	naval computer and telecommunications station
NCW	naval coastal warfare
NCWC	naval coastal warfare commander
NCWS	naval coastal warfare squadron
NDA	national defense area
NDAA	National Defense Authorization Act
NDAF	Navy, Defense Logistics Agency, Air Force
N-day	day an active duty unit is notified for deployment or redeployment
NDB	nondirectional beacon
NDCS	national drug control strategy
NDDOC	US Northern Command Deployment and Distribution Operations Center
NDHQ	National Defence Headquarters, Canada
NDIC	National Drug Intelligence Center
NDL	national desired ground zero list
NDMC	North Atlantic Treaty Organization (NATO) Defense Manpower Committee
NDMS	National Disaster Medical System
NDOC	National Defense Operations Center
NDP	national disclosure policy
NDPB	National Drug Policy Board
NDPC	National Disclosure Policy Committee
NDRF	National Defense Reserve Fleet
NDS	national defense strategy
NDSF	National Defense Sealift Fund
NDU	National Defense University
NEA	Northeast Asia

NEAT	naval embarked advisory team
NEMT	National Emergency Management Team
NEO	noncombatant evacuation operation
NEOCC	noncombatant evacuation operation coordination center
NEPA	National Environmental Policy Act
NEREP	Nuclear Execution and Reporting Plan
NES	National Exploitation System
NESDIS	National Environmental Satellite, Data and Information Service (DOC)
NEST	nuclear emergency support team
NETOPS	network operations
NETS	Nationwide Emergency Telecommunications System
NETT	new equipment training team
NEW	naval expeditionary warfare; net explosive weight
NEWAC	North Atlantic Treaty Organization (NATO) Electronic Warfare Advisory Committee
NEWCS	NATO electronic warfare core staff
NEXCOM	Navy Exchange Command
NFA	no-fire area
NFD	nodal fault diagnostics
NFELC	Naval Facilities Expeditionary Logistics Center
NFESC	Naval Facilities Engineering Service Center
NFI	national foreign intelligence
NFIB	National Foreign Intelligence Board
NFIP	National Flood Insurance Program (FEMA); National Foreign Intelligence Program
NFLIR	navigation forward-looking infrared
NFLS	naval forward logistic site
NFN	national file number
NFO	naval flight officer
NG	National Guard
NGA	National Geospatial-Intelligence Agency
NGB	National Guard Bureau
NGB-OC	National Guard Bureau — Office of the Chaplain
NGF	naval gun fire
NGFS	naval gunfire support
NGIC	National Ground Intelligence Center
NGLO	naval gunfire liaison officer
NGO	nongovernmental organization
NGP	National Geospatial-Intelligence Agency Program
NGRF	National Guard reaction force
NHCS	nonhostile casualty
NI	national identification (number); noted item
NIBRS	National Incident-Based Reporting System
NIC	National Intelligence Council; naval intelligence center
NICCP	National Interdiction Command and Control Plan
NICI	National Interagency Counternarcotics Institute
NID	naval intelligence database
NIDMS	National Military Command System (NMCS) Information for Decision Makers System

NIDS	National Military Command Center (NMCC) information display system
NIE	national intelligence estimate
NIEX	no-notice interoperability exercise
NIEXPG	No-Notice Interoperability Exercise Planning Group
NIFC	national interagency fire center
NII	national information infrastructure
NIIB	National Geospatial Intelligence Agency imagery intelligence brief
NIL	National Information Library
NIMCAMP	National Information Management and Communications Master Plan
NIMS	National Incident Management System
NIP	National Intelligence Program
NIPRNET	Non-Secure Internet Protocol Router Network
NIPS	Naval Intelligence Processing System
NIRT	Nuclear Incident Response Team
NISH	noncombatant evacuation operation (NEO) intelligence support handbook
NISP	national intelligence support plan; Nuclear Weapons Intelligence Support Plan
NIST	National Institute of Standards and Technology; National intelligence support team
NITES	Navy Integrated Tactical Environmental System
NITF	national imagery transmission format
NIU	North Atlantic Treaty Organization (NATO) interface unit
NIWA	naval information warfare activity
NL	Navy lighterage
NLO	naval liaison officer
NL	not less than
NLT	not later than
NLW	nonlethal weapon
NM	network management
nm	nautical mile
NMAWC	Naval Mine and Anti-Submarine Warfare Command
NMB	North Atlantic Treaty Organization (NATO) military body
NMCB	naval mobile construction battalion
NMCC	National Military Command Center
NMCM	not mission capable, maintenance
NMCS	National Military Command System; not mission capable, supply
NMD	national missile defense
NMEC	National Media Exploitation Center
NMET	naval mobile environmental team
NMFS	National Marine Fisheries Services
NMIC	National Maritime Intelligence Center
NMIST	National Military Intelligence Support Team (DIA)
NMOC	network management operations center
NMOSW	Naval METOC Operational Support Web
NMP	national media pool
NMPS	Navy mobilization processing site
NMR	news media representative
NMRC	Naval Medical Research Center

NMS	national military strategy
NMSA	North Atlantic Treaty Organization (NATO) Mutual Support Act
NMS-CO	National Military Strategy for Cyberspace Operations
NNAG	North Atlantic Treaty Organization (NATO) Naval Armaments Group
NOAA	National Oceanic and Atmospheric Administration
NOACT	Navy overseas air cargo terminal
NOC	National Operations Center; network operations center
NOCONTRACT	not releasable to contractors or consultants
NODDS	Naval Oceanographic Data Distribution System
NOE	nap-of-the-earth
NOEA	nuclear operations emergency action
NOFORN	not releasable to foreign nationals
NOG	Nuclear Operations Group
NOGAPS	Navy Operational Global Atmospheric Prediction System
NOHD	nominal ocular hazard distance
NOIC	Naval Operational Intelligence Center
NOK	next of kin
NOMS	Nuclear Operations Monitoring System
NOP	nuclear operations
NOPLAN	no operation plan available or prepared
NORAD	North American Aerospace Defense Command
NORM	normal; not operationally ready, maintenance
NORS	not operationally ready, supply
NOSC	network operations and security center
NOTAM	notice to airmen
NOTMAR	notice to mariners
NP	nonproliferation
NPC	Nonproliferation Center
NPES	Nuclear Planning and Execution System
NPG	nonunit personnel generator
NPS	National Park Service; nonprior service; Nuclear Planning System Navy Postgraduate School
NPT	national pipe thread
NPWIC	National Prisoner of War Information Center
NQ	nonquota
NR	North Atlantic Treaty Organization (NATO) restricted; number
NRC	National Response Center; non-unit-related cargo
NRCC	national response coordination center
NRCHB	Naval Reserve cargo handling battalion
NRCHF	Naval Reserve cargo handling force
NRCHTB	Naval Reserve cargo handling training battalion
NRFI	not ready for issue
NRG	notional requirements generator
NRL	nuclear weapons (NUWEP) reconnaissance list
NRO	National Reconnaissance Office
NROC	Northern Regional Operations Center (CARIBROC-CBRN)
NRP	National Response Plan; non-unit-related personnel
NRPC	Naval Reserve Personnel Center
NRT	near real time
NRTD	near-real-time dissemination

NRZ	non-return-to-zero
NS	nuclear survivability
NSA	national security act; National Security Agency; National security area; national shipping authority; North Atlantic Treaty Organization (NATO) Standardization Agency
NSA/CSS	National Security Agency/Central Security Service
NSAWC	Naval Strike and Air Warfare Center
NSC	National Security Council
NSC/DC	Deputies Committee of the National Security Council
NSCID	National Security Council intelligence directive
NSC/IWG	National Security Council/Interagency Working Group
NSC/PC	National Security Council/Principals Committee
NSC/PCC	National Security Council Policy Coordinating Committee
NSCS	National Security Council System
NSCTI	Naval Special Clearance Team One
NSD	National Security Directive; National Security Division (FBI)
NSDA	non-self deployment aircraft
NSDD	national security decision directive
NSDM	national security decision memorandum
NSDS-E	Navy Satellite Display System-Enhanced
NSE	national support element; Navy support element
NSEP	national security emergency preparedness
NSF	National Science Foundation
NSFS	naval surface fire support
NSG	National System for Geospatial Intelligence; north-seeking gyro
NSGI	National System for Geospatial Intelligence
NSHS	National Strategy for Homeland Security
NSI	not seriously injured
NSL	no-strike list
NSM	national search and rescue (SAR) manual
NSMS	National Strategy for Maritime Security
NSN	National Stock Number
NSNF	nonstrategic nuclear forces
NSO	non-Single Integrated Operational Plan (SIOP) option
NSOC	National Security Operations Center; National Signals Intelligence (SIGINT) Operations Center; Navy Satellite Operations Center
NSOOC	North Atlantic Treaty Organization (NATO) Staff Officer Orientation Course
NSP	national search and rescue plan
N-Sp/CC	North American Aerospace Defense Command (NORAD)-US Space Command/Command Center
NSPD	national security Presidential directive
NSRL	national signals intelligence (SIGINT) requirements list
NSS	National Search and Rescue Supplement; National Security Strategy; national security system; non-self-sustaining
NSSA	National Security Space Architect
NSSCS	non-self-sustaining containership
NSSE	national special security event
NSST	naval space support team
NST	National Geospatial-Intelligence Agency support team

NSTAC	National Security Telecommunications Advisory Committee
NSTISSC	National Security Telecommunications and Information Systems Security Committee
NSTL	national strategic targets list
NSTS	National Secure Telephone System
NSW	naval special warfare
NSWCOM	Naval Special Warfare Command
NSWG	naval special warfare group
NSWTE	naval special warfare task element
NSWTF	naval special warfare task force
NSWTG	naval special warfare task group
NSWTU	naval special warfare task unit
NSWU	naval special warfare unit
NT	nodal terminal
NTACS	Navy tactical air control system
NTAP	National Track Analysis Program
NTB	national target base
NTBC	National Military Joint Intelligence Center Targeting and Battle Damage Assessment Cell
NTC	National Training Center
NTCS-A	Navy Tactical Command System Afloat
NTDS	naval tactical data system
NTF	nuclear task force
N-TFS	New Tactical Forecast System
NTIC	Navy Tactical Intelligence Center
NTISS	National Telecommunications and Information Security System
NTISSI	National Telecommunications and Information Security System (NTISS) Instruction
NTISSP	National Telecommunications and Information Security System (NTISS) Policy
NTMPDE	National Telecommunications Master Plan for Drug Enforcement
NTMS	national telecommunications management structure
NTPS	near-term pre-positioned ships
NTS	night targeting system; noncombatant evacuation operation tracking system
NTSB	National Transportation Safety Board
NTSS	National Time-Sensitive System
NTTP	Navy tactics, techniques, and procedures
NTU	new threat upgrade
NUC	non-unit-related cargo; nuclear
NUCINT	nuclear intelligence
NUDET	nuclear detonation
NUDETS	nuclear detonation detection and reporting system
NUFEA	Navy-unique fleet essential aircraft
NUP	non-unit-related personnel
NURC	non-unit-related cargo
NURP	non-unit-related personnel
NUWEP	policy guidance for the employment of nuclear weapons
NVD	night vision device
NVDT	National Geospatial-Intelligence Agency voluntary deployment team

NVG	night vision goggle
NVS	night vision system
NW	network warfare; not waiverable
NWARS	National Wargaming System
NWB	normal wideband
NWBLTU	normal wideband line termination unit
NWDC	Navy Warfare Development Command
NWFP	Northwest Frontier Province (Pakistan)
NWP	Naval warfare publication; umerical weather prediction
NWREP	nuclear weapons report
NWS	National Weather Service
NWT	normal wideband terminal
O	contour pattern
O&I	operations and intelligence
O&M	operation and maintenance
OA	objective area; operating assembly; operational area; Operations Aerology shipboard METOC division
OADR	originating agency's determination required
OAE	operational area evaluation
OAF	Operation ALLIED FORCE
OAFME	Office of the Armed Forces Medical Examiner
OAG	operations advisory group
OAJCG	Operation Alliance joint control group
OAP	offset aimpoint
OAR	Chairman of the Joint Chiefs of Staff operation plans assessment report
OAS	offensive air support; Organization of American States
OASD	Office of the Assistant Secretary of Defense
OASD(PA)	Office of the Assistant Secretary of Defense (Public Affairs)
OAU	Organization of African Unity
O/B	outboard
OB	operating base; order of battle
OBA	oxygen breathing apparatus
OBFS	offshore bulk fuel system
OBST	obstacle
OBSTINTEL	obstacle intelligence
OC	operations center
OCA	offensive counterair; operational control authority
OCC	Operations Computer Center (USCG)
OCCA	Ocean Cargo Clearance Authority
OCD	orderwire clock distributor
OCDEFT	organized crime drug enforcement task force
OCE	officer conducting the exercise
OCEANCON	control of oceanographic information
OCHA	Office for the Coordination of Humanitarian Affairs
OCJCS	Office of the Chairman of the Joint Chiefs of Staff
OCJCS-PA	Office of the Chairman of the Joint Chiefs of Staff-Public Affairs
OCMI	officer in charge, Marine inspection
OCO	offload control officer
OCONUS	outside the continental United States

OCOP	outline contingency operation plan
OCP	operational configuration processing
OCR	Office of Collateral Responsibility
OCU	orderwire control unit (Types I, II, and III)
OCU-1	orderwire control unit-1
OD	operational detachment; other detainee
ODA	operational detachment-Alpha
ODATE	organization date
O-Day	off-load day
ODB	operational detachment-Bravo
ODC	Office of Defense Cooperation
ODCSLOG	Office of the Deputy Chief of Staff for Logistics (Army)
ODCSOPS	Office of the Deputy Chief of Staff for Operations and Plans (Army)
ODC	Operational Digital Network
ODJS	Office of the Director, Joint Staff
ODR	Office of Defense representative
ODZ	outer defense zone
OEBGD	Overseas Environmental Baseline Guidance Document
OEF	Operation ENDURING FREEDOM
OEG	operational exposure guide; operations security (OPSEC) executive group
OEH	occupational and environmental health
OEM	original equipment manufacturer
OER	officer evaluation report; operational electronic intelligence (ELINT) requirements
OES	office of emergency services
OET	Office of Emergency Transportation (DOT)
OF	officer (NATO)
OFCO	offensive counterintelligence operation
OFDA	Office of US Foreign Disaster Assistance
OFHIS	operational fleet hospital information system
OFOESA	Office of Field Operational and External Support Activities
OGA	other government agency
OGS	overseas ground station
OH	overhead
OI	Office of Intelligence (USCS); operating instruction
OI&A	Office of Intelligence and Analysis (DHS)
OIC	officer in charge
OICC	officer in charge of construction; operational intelligence coordination center
OID	operation order (OPORD) identification
OIF	Operation IRAQI FREEDOM
OIR	operational intelligence requirements; other intelligence requirements
OJT	on-the-job training
OL	operating location
OLD	on-line tests and diagnostics
OLS	operational linescan system; optical landing system
OM	contour multiunit
OMA	Office of Military Affairs (CIA)

OMB	Office of Management and Budget; operations management branch
OMC	Office of Military Cooperation; optical memory card
OMF	officer master file
OMS	Office of Mission Support
OMT	operations management team; orthogonal mode transducer
OMT/OMTP	operational maintenance test(ing)/test plan
ONDCP	Office of National Drug Control Policy
ONE	Operation NOBLE EAGLE
ONI	Office of Naval Intelligence
OOB	order of battle
OOD	officer of the deck
OODA	observe, orient, decide, act
OOS	out of service
OP	observation post; operational publication (USN); ordnance pamphlet
OP3	overt peacetime psychological operations (PSYOP) program
OPARS	Optimum Path Aircraft Routing System
OPBAT	Operation Bahamas, Turks, and Caicos
OPCEN	operations center (USCG)
OPCOM	operational command (NATO)
OPCON	operational control
OPDEC	operational deception
OPDOC	operational documentation
OPDS	offshore petroleum discharge system
OPELINT	operational electronic intelligence
OPFOR	opposing force; opposition force
OPG	operations planning group
OPGEN	operation general matter
OPLAN	operation plan
OPLAW	operational law
OPM	Office of Personnel Management; operations per minute
OPNAVINST	Chief of Naval Operations instruction
OPORD	operation order
OPP	off-load preparation party; orderwire patch panel
OPR	office of primary responsibility
OPREP	operational report
OPS	operational project stock; operations; operations center
OPSCOM	Operations Committee
OPSDEPS	Service Operations Deputies
OPSEC	operations security
OPSTK	operational stock
OPSUM	operation summary
OPT	operational planning team
OPTAR	operating target
OPTASK	operation task
OPTASKLINK	operations task link
OPTEMPO	operating tempo
OPTINT	optical intelligence
OPZONE	operation zone

OR	operational readiness; other rank(s) (NATO)
ORBAT	order of battle
ORCON	originator controlled
ORD	Operational Requirements Document
ORDREF	order reference
ORDTYP	order type
ORG	origin (GEOLOC)
ORIG	origin
ORM	operational risk management
ORP	ocean reception point
OS	operating system
OSA	operational support airlift
OSAT	out-of-service analog test
OSC	on-scene commander; operational support command; operations support center
OSCE	Organization for Security and Cooperation in Europe
OSD	Office of the Secretary of Defense
OSE	on scene endurance; operations support element
OSG	operational support group
OSI	open system interconnection; operational subsystem interface
OSIA	on-site inspection activity
OSINT	open-source intelligence
OSIS	open-source information system
OSO	operational support office
OSOCC	on-site operations coordination center
OSP	operations support package
OSPG	overseas security policy group
OSRI	originating station routing indicator
OSV	ocean station vessel
OT	operational test
OT&E	operational test and evaluation
OTC	officer in tactical command; over the counter
OTG	operational target graphic
OTH	other; over the horizon
OTH-B	over-the-horizon backscatter (radar)
OTHT	over-the-horizon targeting
OTI	Office of Transition Initiatives
OTS	Officer Training School; one-time source
OUB	offshore petroleum discharge system (OPDS) utility boat
OUSD	Office of the Under Secretary of Defense
OUSD(AT&L)	Office of the Under Secretary of Defense (Acquisition, Technology, and Logistics)
OUSD(C)	Office of the Under Secretary of Defense (Comptroller)
OUSD(P)	Office of the Under Secretary of Defense (Policy)
OUT	outsize cargo
OVE	on-vehicle equipment
OVER	oversize cargo
OVM	Operation Vigilant Mariner
OW	orderwire

P	parallel pattern; priority
PA	parent relay; physician assistant; primary agency; probability of arrival; public affairs
PAA	primary aircraft authorization
PABX	private automatic branch exchange (telephone)
PACAF	Pacific Air Forces
PAD	patient administration director; positional adjustment; precision aircraft direction
PADD	person authorized to direct disposition of human remains
PADS	position azimuth determining system
PAG	public affairs guidance
PAI	primary aircraft inventory
PAL	permissive action link; personnel allowance list; program assembler language
PALS	precision approach landing system
PAM	pulse amplitude modulation
PaM	passage material
PANS	procedures for air navigation services
PAO	public affairs office; public affairs officer
PAR	performance assessment report; population at risk; precision approach radar
PARC	principal assistant for contracting
PARKHILL	high frequency cryptological device
PARPRO	peacetime application of reconnaissance programs
PAS	personnel accounting symbol
PAT	public affairs team
PAV	policy assessment visit
PAWS	phased array warning system
PAX	passengers; public affairs plans
PB	particle beam; patrol boat; peace building; President's budget
PBA	performance-based agreement; production base analysis
PBCR	portable bar code recorder
PBD	program budget decision
PC	patrol craft; personal computer; pilot in command; preliminary coordination; Principals Committee
Pc	cumulative probability of detection
P,C,&H	packing, crating, and handling
PCA	Posse Comitatus Act
PCC	policy coordination committee; primary control center
PCF	personnel control facility
PCL	positive control launch
PC-LITE	processor, laptop imagery transmission equipment
PCM	pulse code modulation
PCO	primary control officer; procuring contracting officer
PCRTS	primary casualty receiving and treatment ship
PCS	permanent change of station; personal communications system; primary control ship; processing subsystem; processor controlled strapping

PCT	personnel control team
PCTC	pure car and truck carrier
PCZ	physical control zone
PD	position description; Presidential directive; probability of damage; probability of detection; procedures description; program definition; program directive; program director; public diplomacy
Pd	drift compensated parallelogram pattern
PDA	preliminary damage assessment
PDAI	primary development/test aircraft inventory
PDD	Presidential decision directive
PDDA	power driven decontamination apparatus
PDDG	program directive development group
PDG	positional data graphic
PDM	program decision memorandum
PDOP	position dilution of precision
PDS	position determining system; primary distribution site; protected distribution system
PDUSD(P&R)	Principal Deputy Under Secretary of Defense (Personnel & Readiness)
PE	peace enforcement; peacetime establishment; personal effects; program element
PEAD	Presidential emergency action document
PEAS	psychological operations (PSYOP) effects analysis subsystem
PEC	program element code
PECK	patient evacuation contingency kit
PECP	precision engagement collaboration process
PEDB	planning and execution database
PEGEO	personnel geographic location
PEI	principal end item
PEM	program element monitor
PEO	peace enforcement operations
PEP	personnel exchange program
PER	personnel
PERE	person eligible to receive effects
PERID	period
PERINTSUM	periodic intelligence summary
PERMREP	permanent representative (NATO)
PERSCO	personnel support for contingency operations
PERSCOM	personnel command (Army)
PERSINS	personnel information system
PES	preparedness evaluation system
PFA	primary federal agency
PFD	personal flotation device
PFDB	planning factors database
PFIAB	President's Foreign Intelligence Advisory Board
PFID	positive friendly identification
PFO	principal federal official
PFP	Partnership for Peace (NATO)
PGM	precision-guided munition
pH	potential of hydrogen
PHIBCB	amphibious construction battalion

PHIBGRU	amphibious group
PHIBOP	amphibious operation
PHIBRON	amphibious squadron
PHO	posthostilities operations
PHOTINT	photographic intelligence
PHS	Public Health Service
PI	point of impact; probability of incapacitation; procedural item
PIC	parent indicator code; payment in cash; person identification code; pilot in command; press information center (NATO)
PID	plan identification number
PIDD	planned inactivation or discontinued date
PIF	problem identification flag
PII	pre-incident indicators
PIM	pretrained individual manpower
PIN	personnel increment number
PINS	precise integrated navigation system
PIO	press information officer; public information officer
PIPS	plans integration partitioning system
PIR	priority intelligence requirement
PIRAZ	positive identification and radar advisory zone
PIREP	pilot report
PIW	person in water
PJ	pararescue jumper
PK	peacekeeping; probability of kill
PKG-POL	packaged petroleum, oils, and lubricants
PKI	public key infrastructure
PKO	peacekeeping operations
PKP	purple k powder
PL	phase line; public law
PLA	plain language address
PLAD	plain language address directory
PLANORD	planning order
PLAT	pilot's landing aid television
PLB	personal locator beacon
PLC	power line conditioner
PLGR	precise lightweight global positioning system (GPS) receiver
PLL	phase locked loop
PLL/ASL	prescribed load list/authorized stock level
PLRS	position location reporting system
PLS	palletized load system; personal locator system; personnel locator system; pillars of logistic support; precision location system
PLT	platoon; program library tape
PM	parallel track multiunit passage material; patient movement; peacemaking; political-military affairs; preventive medicine; program manager; provost marshal
PMA	political/military assessment
PMAA	Production Management Alternative Architecture
PMAI	primary mission aircraft inventory
P/M/C	passengers/mail/cargo
PMC	parallel multiunit circle; partial mission-capable

PMCM	partial mission-capable, maintenance
PMCS	partial mission-capable, supply
PMCT	port movement control team
PMD	program management directive
PME	professional military education
PMEL	precision measurement equipment laboratory
PMGM	program manager's guidance memorandum
PMI	patient movement item
PMIS	psychological operations (PSYOP) management information subsystem
PMN	parallel track multiunit non-return
PMO	production management office(r); program management office
PMOS	primary military occupational specialty
PMR	parallel track multiunit return; patient movement request; patient movement requirement
PMRC	patient movement requirements center
PMS	portable meteorological subsystem
PN	partner nation; pseudonoise
PNID	precedence network in dialing
PNVS	pilot night vision system
P/O	part of
PO	peace operations; petty officer
POA	plan of action
POADS	psychological operations automated data system
POAI	primary other aircraft inventory
POAS	psychological operations automated system
POAT	psychological operations assessment team
POB	persons on board; psychological operations battalion
POC	point of contact
POCD	port operations cargo detachment
POD	plan of the day; port of debarkation; probability of detection
POE	port of embarkation; port of entry
POF	priority of fires
POG	port operations group; psychological operations group
POI	period of interest; program of instruction
POL	petroleum, oils, and lubricants
POLAD	political advisor
POLCAP	bulk petroleum capabilities report
POLMIL	political-military
POM	program objective memorandum
POMCUS	pre-positioning of materiel configured to unit sets
POMSO	Plans, Operations, and Military Support Office(r) (NG)
POP	performance oriented packaging
POPS	port operational performance simulator
POR	proposed operational requirement
PORTS	portable remote telecommunications system
PORTSIM	port simulation model
POS	peacetime operating stocks; Point of Sale; probability of success
POSF	port of support file
POSSUB	possible submarine
POSTMOB	post mobilization

POTF	psychological operations task force
POTG	psychological operations task group
POTUS	President of the United States
POV	privately owned vehicle
POW	prisoner of war
P/P	patch panel
p-p	peak-to-peak
PPA	personnel information system (PERSINS) personnel activity
PPAG	proposed public affairs guidance
PPBE	Planning, Programming, Budgeting, and Execution
PPD	program planning document
PPDB	point positioning database
PPE	personal protective equipment
PPF	personnel processing file
Pplan	programming plan
PPLI	precise participant location and identification
ppm	parts per million
PPP	power projection platform; primary patch panel; priority placement program
PPR	prior permission required
PPS	precision positioning service
PR	personnel recovery; Phoenix Raven; primary zone; production requirement; program review
PRA	patient reception area; primary review authority
PRANG	Puerto Rican Air National Guard
PRBS	pseudorandom binary sequence
PRC	populace and resources control; Presidential Reserve Callup
PRCC	personnel recovery coordination cell
PRD	personnel readiness division; Presidential review directive
PRDO	personnel recovery duty officer
PRECOM	preliminary communications search
PREMOB	pre-mobilization
PREPO	pre-positioned force, equipment, or supplies; prepositioning
PREREP	pre-arrival report
PRF	personnel resources file; pulse repetition frequency
PRG	program review group
PRI	movement priority for forces having the same latest arrival date (LAD); priority; progressive routing indicator
PRIFLY	primary flight control
Prime BEEF	Prime Base Engineer Emergency Force
PRISM	Planning Tool for Resource, Integration, Synchronization, and Management
PRM	Presidential review memorandum
PRMFL	perm file
PRMS	personnel recovery mission software
PRN	pseudorandom noise
PRO	personnel recovery officer
PROBSUB	probable submarine
PROC	processor; Puerto Rican Operations Center
PROFIS	professional officer filler information system

PROM	programmable read-only memory
PROPIN	caution - proprietary information involved
PROVORG	providing organization
proword	procedure word
PRP	personnel reliability program
PRRIS	Puerto Rican radar integration system
PRSL	primary zone/switch location
PRT	pararescue team; patient reception team
PRTF	personnel recovery task force
PRU	pararescue unit; primary reporting unit
PS	parallel track single-unit; processing subsystem
PSA	port support activity
PSB	poststrike base
PSC	port security company; principal subordinate command
PSD	planning systems division
PSE	peculiar support equipment; psychological operations support element
PS/HD	port security/harbor defense
PSHDGRU	port security and harbor defense group
PSI	personnel security investigation; Proliferation Security Initiative
psi	pounds per square inch
PSK	phase shift keying
PSL	parallel track single-unit long-range aid to navigation (LORAN)
PSMS	Personnel Status Monitoring System
PSN	packet switching node; public switch network
PSO	peace support operations (NATO); post security officer
PSP	perforated steel planking; portable sensor platform; power support platform
PSPS	psychological operations (PSYOP) studies program subsystem
PSS	parallel single-unit spiral
P-STATIC	precipitation static
PSTN	public switched telephone network
PSU	port security unit
PSV	pseudosynthetic video
PSYOP	psychological operations
PTA	position, time, altitude
PTAI	primary training aircraft inventory
PTC	peace through confrontation; primary traffic channel
PTDO	prepare to deploy order
PTT	postal telephone and telegraph; public telephone and telegraph; push-to-talk
PTTI	precise time and time interval
pub	publication
PUK	packup kit
PUL	parent unit level
PV	prime vendor
PVNTMED	preventive medicine
PVT	positioning, velocity, and timing
PW	prisoner of war
pW	picowatt
PWB	printed wiring board (assembly)

PWD	programmed warhead detonation
PWF	personnel working file
PWIS	Prisoner of War Information System
PWR	pre-positioned wartime reserves
PWRMR	pre-positioned war materiel requirement
PWRMS	pre-positioned war reserve materiel stock
PWRR	petroleum war reserve requirements
PWRS	petroleum war reserve stocks; pre-positioned war reserve stock
PZ	pickup zone
QA	quality assurance
QAM	quadrature amplitude modulation
QAT	quality assurance team
QC	quality control
QD	quality distance
QDR	quality deficiency report
QEEM	quick erect expandable mast
QHDA	qualified hazardous duty area
QM	quartermaster
QPSK	quadrature phase shift keying
QRA	quick reaction antenna
QRCT	quick reaction communications terminal
QRE	quick reaction element
QRF	quick response force
QRG	quick response graphic
QRP	quick response posture
QRS	quick reaction strike
QRSA	quick reaction satellite antenna
QRT	quick reaction team
QS	quality surveillance
Q-ship	decoy ship
QSR	quality surveillance representative
QSTAG	quadripartite standardization agreement
QTY	quantity
QUADCON	quadruple container
R	routine; search radius
R&D	research and development
R&R	rest and recuperation
R&S	reconnaissance and surveillance
R2P2	rapid response planning process
RA	response action; risk analysis; risk assessment
RAA	redeployment assembly area
RABFAC	radar beacon forward air controller
RAC	responsible analytic center
RAC-OT	readiness assessment system - output tool
RADAREXREP	radar exploitation report
RADAY	radio day
RADBN	radio battalion
RADC	regional air defense commander

RADCON	radiological control team
RADF	radarfind
RADHAZ	electromagnetic radiation hazards
RADINT	radar intelligence
RADS	rapid area distribution support (USAF)
RAE	right of assistance entry
RAF	Royal Air Force (UK)
R-AFF	regimental affiliation
RAM	raised angle marker; random access memory; random antiterrorism measure
RAMCC	regional air movement coordination center
RAOB	rawindsonde observation
RAOC	rear area operations center; regional air operations center
RAP	Radiological Assistance Program (DOE); rear area protection; Remedial Action Projects Program (JCS)
RAS	recovery activation signal; refueling at sea
RAST	recovery assistance, securing, and traversing systems
RASU	random access storage unit
RATT	radio teletype
RB	radar beacon; short-range coastal or river boat
RBC	red blood cell
RBE	remain-behind equipment
RBECS	Revised Battlefield Electronic Communications, Electronics, Intelligence, and Operations (CEIO) System
RBI	RED/BLACK isolator
RB std	rubidium standard
RC	receive clock regional coordinator; Reserve Component; river current
RCA	residual capabilities assessment; riot control agent
RCAT	regional counterdrug analysis team
RCC	relocation coordination center
RCCPDS	Reserve Component common personnel data system
RCEM	regional contingency engineering management
RCHB	reserve cargo handling battalion
RCIED	radio-controlled improvised explosive device
RCM	Rules for Courts-Martial
RCMP	Royal Canadian Mounted Police
RC NORTH	Regional Command North (NATO)
RCO	regional coordinating office (DOE)
RCP	resynchronization control panel
RCS	radar cross section
RC SOUTH	Regional Command South (NATO)
RCSP	remote call service position
RCT	regimental combat team; rescue coordination team (Navy)
RCTA	Regional Counterdrug Training Academy
RCU	rate changes unit; remote control unit
RCVR	receiver
RD	receive data; ringdown
RDA	research, development, and acquisition
R-day	redeployment day
RDCFP	Regional Defense Counterterrorism Fellowship Program

RDD	radiological dispersal device; required delivery date
RDECOM	US Army Research, Development, and Engineering Command
RDF	radio direction finder; rapid deployment force
RDO	request for deployment order
RDT&E	research, development, test and evaluation
REACT	rapid execution and combat targeting
REAC/TS	radiation emergency assistance center/training site (DOE)
READY	resource augmentation duty program
RECA	Residual Capability Assessment
RECAS	residual capability assessment system
RECAT	residual capability assessment team
RECCE	reconnaissance
RECCEXREP	reconnaissance exploitation report
RECMOB	reconstitution-mobilization
RECON	reconnaissance
RED HORSE	Rapid Engineers Deployable Heavy Operations Repair Squadron, Engineers
REF	reference(s)
REGT	regiment
REL	relative
RELCAN	releasable to Canada
REM	roentgen equivalent mammal
REMT	regional emergency management team
REMUS	remote environmental monitoring unit system
REPOL	bulk petroleum contingency report; petroleum damage and deficiency report; reporting emergency petroleum, oils, and lubricants
REPSHIP	report of shipment
REPUNIT	reporting unit
REQCONF	request confirmation
REQSTATASK	air mission request status tasking
RESA	research, evaluation, and system analysis
RESCAP	rescue combat air patrol
RESCORT	rescue escort
RESPROD	responsible production
RET	retired
RF	radio frequency; reserve force; response force
RFA	radio frequency authorization; request for assistance; restrictive fire area
RFC	response force commander
RF CM	radio frequency countermeasures
RFD	revision first draft
RF/EMPINT	radio frequency/electromagnetic pulse intelligence
RFF	request for feedback; request for forces
RFI	radio frequency interference; ready for issue; request for information
RFID	radio frequency identification
RFL	restrictive fire line
RFP	request for proposal
RFS	request for service
RFW	request for waiver
RG	reconstitution group

RGR	Rangers
RGS	remote geospatial intelligence services
RH	reentry home
Rh	Rhesus factor
RHIB	rigid hull inflatable boat
RI	radiation intensity; Refugees International; Routing indicator
RIB	rubberized inflatable boat
RIC	routing indicator code
RICO	regional interface control officer
RIG	recognition identification group
RIK	replacement in kind
RIMS	registrant information management system
RINT	unintentional radiation intelligence
RIP	register of intelligence publications
RIS	reconnaissance information system
RISOP	red integrated strategic offensive plan
RISTA	reconnaissance, intelligence, surveillance, and target acquisition
RIT	remote imagery transceiver
RJTD	reconstitution joint table of distribution
RLD	ready-to-load date
RLE	rail liaison element
RLG	regional liaison group; ring laser gyro
RLGM	remote loop group multiplexer
RLGM/CD	remote loop group multiplexer/cable driver
RLP	remote line printer
RM	recovery mechanism; resource management; risk management
RMC	remote multiplexer combiner; rescue mission commander; Resource Management Committee (CSIF); returned to military control
RMKS	remarks
RMO	regional Marine officer
RMP	religious ministry professional
RMS	requirements management system; root-mean-square
RMU	receiver matrix unit
RNAV	area navigation
RNP	remote network processor
R/O	receive only
Ro	search radius rounded to next highest whole number
ROA	restricted operations area
ROC	regional operations center; required operational capability
ROCU	remote orderwire control unit
ROE	rules of engagement
ROEX	rules of engagement exercise
ROG	railhead operations group
ROICC	resident officer in charge of construction
ROK	Republic of Korea
ROM	read-only memory; rough order of magnitude
RON	remain overnight
RO/RO	roll-on/roll-off
ROS	reduced operating status

ROTC	Reserve Officer Training Corps
ROTHR	relocatable over-the-horizon backscatter radar (USN)
ROWPU	reverse osmosis water purification unit
ROZ	restricted operations zone
RP	reconstitution priority; release point (road); retained personnel
RPPO	Requirements, Plans, and Policy Office
RPT	report
RPTOR	reporting organization
RPV	remotely piloted vehicle
RQMT	requirement
RQT	rapid query tool
RR	reattack recommendation
RRC	regional reporting center
RRCC	regional response coordination center
RRDF	roll-on/roll-off (RO/RO) discharge facility
RRF	rapid reaction force; rapid response force; Ready Reserve Fleet; Ready Reserve Force
RRPP	rapid response planning process
RS	rate synthesizer; religious support; requirement submission
RSA	retrograde storage area
RSC	red station clock; regional service center; rescue sub-center
RSD	reporting of supply discrepancy
RSE	retrograde support element
RSG	reference signal generator
RSI	rationalization, standardization, and interoperability
RSL	received signal level
RSN	role specialist nation
RSO	regional security officer
RSOC	regional signals intelligence (SIGINT) operations center
RSOI	reception, staging, onward movement, and integration
RSP	recognized surface picture; Red Switch Project (DOD); religious support plan; religious support policy
RSPA	Research and Special Programs Administration
RSS	radio subsystem; remote sensors subsystem; root-sum-squared
RSSC	regional satellite communications (SATCOM) support center; regional satellite support cell; regional signals intelligence (SIGINT) support center (NSA); regional space support center
RSSC-LO	regional space support center liaison officer
RST	religious support team
RSTA	reconnaissance, surveillance, and target acquisition
RSTV	real-time synthetic video
RSU	rapid support unit; rear support unit; remote switching unit
R/T	receiver/transmitter
RT	recovery team; remote terminal; rough terrain
RTA	residual threat assessment
RTB	return to base
RTCC	rough terrain container crane
RTCH	rough terrain container handler
RTD	returned to duty
RTF	regional task force; return to force

RTFL	rough terrain forklift
RTG	radar target graphic
RTL	restricted target list
RTLP	receiver test level point
RTM	real-time mode
RTOC	rear tactical operations center
RTS	remote transfer switch
RTTY	radio teletype
RU	release unit; rescue unit
RUF	rules for the use of force
RUIC	Reserve unit identification number
RUSCOM	rapid ultrahigh frequency (UHF) satellite communications
RV	long-range seagoing rescue vessel; reentry vehicle; rekeying variable; rendezvous
RVR	runway visibility recorder
RVT	remote video terminal
RWCM	regional wartime construction manager
RWI	radio and wire integration
RWR	radar warning receiver
RWS	rawinsonde subsystem
RX	receive; receiver
RZ	recovery zone
	return-to-zero
S&F	store-and-forward
S&M	scheduling and movement
S&R	search and recovery
S&T	science and technology; scientific and technical
S&TI	scientific and technical intelligence
S-2	battalion or brigade intelligence staff officer (Army; Marine Corps battalion or regiment)
S-3	battalion or brigade operations staff officer (Army; Marine Corps battalion or regiment)
S-4	battalion or brigade logistics staff officer (Army; Marine Corps battalion or regiment)
SA	security assistance; selective availability (GPS); senior adviser; situational awareness; staging area; stand-alone switch
SAA	senior airfield authority
SAAF	small austere airfield
SAAFR	standard use Army aircraft flight route
SAAM	special assignment airlift mission
SAB	scientific advisory board (USAF)
SABER	situational awareness beacon with reply
SAC	special actions cell; special agent in charge; supporting arms coordinator
SACC	supporting arms coordination center
SACEUR	Supreme Allied Commander, Europe (NATO)
SACLANT	Supreme Allied Command, Atlantic
SACS	secure telephone unit (STU) access control system
SACT	Supreme Allied Commander Transformation

SADC	sector air defense commander
SADL	situation awareness data link
SAF	Secretary of the Air Force
SAFE	secure analyst file environment; selected area for evasion
SAFE-CP	selected area for evasion-contact point
SAFER	evasion and recovery selected area for evasion (SAFE) area activation request
SAFWIN	secure Air Force weather information network
SAG	surface action group
SAI	sea-to-air interface; single agency item
SAL	small arms locker
SAL-GP	semiactive laser-guided projectile (USN)
SALM	single-anchor leg mooring
SALT	supporting arms liaison team
SALTS	streamlined automated logistics transfer system; streamlined automated logistics transmission system
SALUTE	size, activity, location, unit, time, and equipment
SAM	space available mail; special airlift mission; surface-to-air missile
SAMM	security assistance management manual
SAMS	School of Advanced Military Studies
SAO s	Security assistance office/officer; security assistance organization; selected attack option
SAOC	sector air operations center
SAP	special access program
SAPI	special access program for intelligence
SAPO	subarea petroleum office
SAPR	sexual assault prevention and response
SAR	satellite access request; search and rescue; special access requirement; suspicious activity report; synthetic aperture radar
SARC	sexual assault response coordinator; surveillance and reconnaissance center
SARDOT	search and rescue point
SARIR	search and rescue incident report
SARMIS	search and rescue management information system
SARNEG	search and rescue numerical encryption group
SARREQ	search and rescue request
SARSAT	search and rescue satellite-aided tracking
SARSIT	search and rescue situation summary report
SARTEL	search and rescue (SAR) telephone (private hotline)
SARTF	search and rescue task force
SAS	sealed authenticator system; special ammunition storage
SASP	special ammunition supply point
SASS	supporting arms special staff
SASSY	supported activities supply systems
SAT	satellite; security alert team
SATCOM	satellite communications
SAU	search attack unit
SAW	surface acoustic wave
SB	standby base
SBCT	Stryker brigade combat team

SBIRS	space-based infrared system
SBL	space-based laser
SBPO	Service blood program officer
SBR	special boat squadron
SBRPT	subordinate reporting organization
SBS	senior battle staff; support battle staff
SBSS	science-based stockpile stewardship
SBT	special boat team
SBSO	sustainment brigade special operations
SBU	special boat unit
SC	sea current; search and rescue coordinator; station clock; strategic communication
SCA	space coordinating authority
SCAR	strike coordination and reconnaissance
SCAS	stability control augment system
SCATANA	security control of air traffic and navigation aids
SC ATLANTIC	Strategic Command, Atlantic (NATO)
SCATMINE	scatterable mine
SCATMINEWARN	scatterable minefield warning
SCC	security classification code; Space Control Center (USSPACECOM); shipping coordination center; Standards Coordinating Committee
SCDL	surveillance control data link
SCE	Service cryptologic element
SC EUROPE	Strategic Command, Europe (NATO)
SCF(UK)	Save the Children Fund (United Kingdom)
SCF(US)	Save the Children Federation (United States)
SCG	Security Cooperation Guidance; switching controller group
SCI	sensitive compartmented information
SCIF	sensitive compartmented information facility
SCL	standard conventional load
SCM	security countermeasure
SCMP	strategic command, control, and communications (C3) master plan
SCNE	self-contained navigation equipment
SCO	senior contracting official; state coordinating officer
SCOC	systems control and operations concept
SCONUM	ship control number
SCP	secure conferencing project; security cooperation plan; system change proposal
SCPT	strategic connectivity performance test
SCRB	software configuration review board
SCT	shipping coordination team; single channel transponder
SCTIS	single channel transponder injection system
SCTS	single channel transponder system
SCT-UR	single channel transponder ultrahigh frequency (UHF) receiver
SCUD	surface-to-surface missile system
SDA	Seventh-Day Adventist (ADRA)
S-day	day the President authorizes selective reserve call-up
SDB	Satellite Communications Database
SDDC	Surface Deployment and Distribution Command

SDDCTEA	Surface Deployment and Distribution Command Transportation Engineering Agency
SDF	self defense force
SDIO	Strategic Defense Initiative Organization
SDLS	satellite data link standards
SDMX	space division matrix
SDN	system development notification
SDNRIU	secure digital net radio interface unit
SDO	ship's debarkation officer
SDP	strategic distribution platform
SDR	system design review
SDSG	space division switching group
SDSM	space division switching matrix
SDV	sea-air-land team (SEAL) delivery vehicle; Submerged delivery vehicle
SEA	Southeast Asia
SEABEE	Navy construction engineer; sea barge
SEAD	suppression of enemy air defenses
SEAL	sea-air-land team
SEAVAN	military container moved via ocean
SEC	submarine element coordinator
SECAF	Secretary of the Air Force
SECARMY	Secretary of the Army
SecDef	Secretary of Defense
SECDHS	Secretary of the Department of Homeland Security
SECNAV	Secretary of the Navy
SECNAVINST	Secretary of the Navy instruction
SECOMP	secure en route communications package
SECORD	secure cord switchboard
SECRA	secondary radar data only
SECSTATE	Secretary of State
SECTRANS	Secretary of Transportation
SED	signals external data
SEDAS	spurious emission detection acquisition system
SEF	sealift enhancement feature
SEHS	special events for homeland security
SEI	specific emitter identification
SEL REL	selective release
SELRES	Selected Reserve
SEMA	special electronic mission aircraft
SEMS	standard embarkation management system
SEO/SEP	special enforcement operation/special enforcement program
SEP	signal entrance panel; spherical error probable
SEPLO	state emergency preparedness liaison officer
SERE	survival, evasion, resistance, and escape
SERER	survival, evasion, resistance, escape, recovery
SES	senior executive service
SETA	system engineering and technical assistance
SEW	shared early warning
S/EWCC	signals intelligence/electronic warfare coordination center
SEWG	Special Events Working Group

S/EWOC	signals intelligence/electronic warfare operations center
SEWS	satellite early warning system
SF	security force; security forces (Air Force or Navy); single frequency; special forces; standard form
SFAF	standard frequency action format
SFCP	shore fire control party
SFG	security forces group; special forces group
SFI	spectral composition
SFLEO	senior federal law enforcement official
SFMS	special forces medical sergeant
SFOB	special forces operations base
SFOD-A/B/C	special forces operational detachment-A/B/C
SFOR	Stabilization Force
SFS	security forces squadron
SG	strike group; supergroup; Surgeon General
SGEMP	system-generated electromagnetic pulse
SGSA	squadron group systems advisor
SGSI	stabilized glide slope indicator
SHAPE	Supreme Headquarters Allied Powers, Europe
SHD	special handling designator
SHF	super-high frequency
SHORAD	short-range air defense
SHORADEZ	short-range air defense engagement zone
SI	special intelligence
SIA	station of initial assignment
SIAGL	survey instrument azimuth gyroscope lightweight
SIC	subject identification code
SICO	sector interface control officer
SICR	specific intelligence collection requirement
SID	secondary imagery dissemination; standard instrument departure
SIDAC	single integrated damage analysis capability
SIDL	standard intelligence documents list
SIDS	secondary imagery dissemination system
SIF	selective identification feature
SIG	signal
SIGINT	signals intelligence
SIGSEC	signal security
SII	seriously ill or injured; statement of intelligence interest
SIM	system impact message
SIMLM	single integrated medical logistics management; Single integrated medical logistics manager
SINCGARS	single-channel ground and airborne radio system
SINS	ship's inertial navigation system
SIO	senior intelligence officer; special information operations
SIOP	Single Integrated Operational Plan
SIOP-ESI	Single Integrated Operational Plan-Extremely Sensitive Information
SIPRNET	SECRET Internet Protocol Router Network
SIR	serious incident report; specific information requirement
SIRADS	stored imagery repository and dissemination system
SIRMO	senior information resources management official

SIS	special information systems
SITLM	single integrated theater logistic manager
SITREP	situation report
SIV	special interest vessel
SJA	staff judge advocate
SJFHQ	standing joint force headquarters
SJFHQ(CE)	standing joint force headquarters (core element)
SJFHQ-N	Standing Joint Force Headquarters - North
SJS	Secretary, Joint Staff
SKE	station-keeping equipment
SL	sea level; switch locator
SLA	service level agreement
SLAM	stand-off land attack missile
SLAR	side-looking airborne radar
SLBM	submarine-launched ballistic missile
SLC	satellite laser communications; single line concept
SLCM	sea-launched cruise missile
SLCP	ship lighterage control point; ship's loading characteristics pamphlet
SLD	system link designator
SLEP	service life extension program
SLGR	small, lightweight ground receiver (GPS)
SLIT	serial-lot item tracking
SLO	space liaison officer
SLOC	sea line of communications
SLP	seaward launch point
SLRP	survey, liaison, and reconnaissance party
SLWT	side loadable warping tug
SM	Secretary, Joint Staff, memorandum; Service manager; Staff memorandum; system manager
SMART	special medical augmentation response team
SMART-AIT	special medical augmentation response – aeromedical isolation team
SMC	midpoint compromise track spacing; search and rescue mission coordinator; system master catalog
SMCA	single manager for conventional ammunition
SMCC	strategic mobile command center
SMCM	surface mine countermeasures
SMCOO	spectrum management concept of operations
SMCR	Selected Marine Corps Reserve
SMD	strategic missile defense
SMDC	Space & Missile Defense Command (Army)
SME	subject matter expert
SMEB	significant military exercise brief
SMEO	small end office
SMFT	semi-trailer mounted fabric tank
SMI	security management infrastructure
SMIO	search and rescue (SAR) mission information officer
SMO	senior meteorological and oceanographic officer; strategic mobility office(r); support to military operations
SMP	sub-motor pool
SMPT	School of Military Packaging Technology

SMRI	service message routing indicator
SMS	single mobility system
SMTP	simple message transfer protocol
SMU	special mission unit; supported activities supply system (SASSY) management unit
S/N	signal to noise
SNCO	staff noncommissioned officer
SNF	strategic nuclear forces
SNIE	special national intelligence estimates
SNLC	Senior North Atlantic Treaty Organization (NATO) Logisticians Conference
SNM	system notification message
SNOI	signal not of interest
SO	safety observer; special operations
SOA	separate operating agency; special operations aviation; speed of advance; status of action; sustained operations ashore
SOAF	status of action file
SOC	security operations center; special operations commander
SOCA	special operations communications assembly
SOCC	Sector Operations Control Center (NORAD)
SOCCE	special operations command and control element
SOCCENT	Special Operations Component, United States Central Command
SOCCET	special operations critical care evacuation team
SOCCT	special operations combat control team
SOCEUR	Special Operations Component, United States European Command
SOCEX	special operations capable exercise
SOCJFCOM	Special Operations Command, Joint Forces Command
SOCOORD	special operations coordination element
SOCP	special operations communication package
SOCPAC	Special Operations Component, United States Pacific Command
SOCRATES	Special Operations Command, Research, Analysis, and Threat Evaluation System
SOCSOUTH	Special Operations Component, United States Southern Command
SOD	special operations division; strategy and options decision (Planning, Programming, and Budgeting System)
SODARS	special operations debrief and retrieval system
SOE	special operations executive
SOF	special operations forces; supervisor of flying
SOFA	status-of-forces agreement
SOFAR	sound fixing and ranging
SOFLAM	special operations laser marker
SOFME	special operations forces medical element
SOFSA	special operations forces support activity
SOG	special operations group
SOI	signal of interest; signal operating instructions; space object identification
SOIC	senior officer of the Intelligence Community
SOLAS	safety of life at sea
SOLE	special operations liaison element
SOLIS	signals intelligence (SIGINT) On-line Information System

SOLL	special operations low-level
SOM	satellite communications operational manager; start of message; system operational manager
SOMA	status of mission agreement
SOMARDS	Standard Operation and Maintenance Army Research and Development System
SOMARDS NT	Standard Operation and Maintenance Army Research and Development System Non-Technical
SOMPF	special operations mission planning folder
SONMET	special operations naval mobile environment team SoO ship of opportunity
SOOP	Center for Operations, Plans, and Policy
SOP	standard operating procedure; standing operating procedure
SOPA	senior officer present afloat (USN)
SO-peculiar	special operations-peculiar
SOR	statement of requirement
SORTIEALOT	sortie allotment message
SORTS	Status of Resources and Training System
SOS	special operations squadron
SOSB	special operations support battalion
SOSC	special operations support command (theater army)
SOSCOM	special operations support command
SOSE	special operations staff element
SOSG	station operations support group
SOSR	suppress, obscure, secure, and reduce
SOTA	signals intelligence (SIGINT) operational tasking authority
SOTAC	special operations terminal attack controller
SOTF	special operations task force
SOTSE	special operations theater support element
SOUTHAF	Southern Command Air Forces
SOUTHROC	Southern Region Operational Center (USSOUTHCOM)
SOW	special operations wing; standoff weapon; statement of work
SOWT	special operations weather team
SOWT/TE	special operations weather team/tactical element
SP	security police
SPA	special psychological operations (PSYOP) assessment; submarine patrol area
SPACEAF	Space Air Forces
SPACECON	control of space information
SPCC	ships parts control center (USN)
SPEAR	strike protection evaluation and antiair research
SPEC	specified
SPECAT	special category
SPECWAR	special warfare
SPG	Strategic Planning Guidance
SPI	special investigative (USAF)
SPINS	special instructions
SPINTCOMM	special intelligence communications handling system
SPIREP	specialist intelligence report; spot intelligence report
SPLX	simplex

SPM	single point mooring; single port manager
SPMAGTF	special purpose Marine air-ground task force
SPO	system program office
SPOC	search and rescue (SAR) points of contact; space command operations center; Space Operations Center (USSPACECOM)
SPOD	seaport of debarkation
SPOE	seaport of embarkation
SPOTREP	spot report
SPP	Security and Prosperity Partnership of North America; Shared production program
SPR	software problem report
SPRINT	special psychiatric rapid intervention team
SPS	special psychological operations (PSYOP) study; standard positioning system
SPSC	system planning and system control
SPTCONF	support confirmation
SPTD CMD	supported command
SPTG CMD	supporting command
SPTREQ	support request
sqft	square feet
SR	special reconnaissance
SRA	specialized-repair activity
SRAM	short-range air-to-surface attack missile; System replacement and modernization
SRB	software release bulletin; system review board (JOPES)
SRBM	short-range ballistic missile
SRC	security risk category; service reception center; Single Integrated Operational Plan (SIOP) response cell; standard requirements code; survival recovery center
SRCC	service reserve coordination center
SRF	secure Reserve force
SRG	Seabee readiness group; short-range aircraft
SRI	surveillance, reconnaissance, and intelligence (Marine Corps)
SRIG	surveillance, reconnaissance, and intelligence group (USMC)
SROC	Senior Readiness Oversight Council; Southern Region Operational Center, United States Southern Command
SROE	standing rules of engagement
SRP	Sealift Readiness Program; sealift reserve program; seaward recovery point; Single Integrated Operational Plan (SIOP) reconnaissance plan
SRP/PDS	stabilization reference package/position determining system
SRR	search and rescue region
SRS	search and rescue sector
SRSG	special representative of the Secretary-General
SRT	scheduled return time; special reaction team; standard remote terminal; strategic relocatable target
SRTD	signals research and target development
S/RTF	search and recovery task force
SRU	search and rescue unit
SR-UAV	short-range unmanned aerial vehicle
SRUF	standing rules for the use of force

SRWBR	short range wide band radio
S/S	steamship
SS	submarine
SSA	software support activity; special support activity (NSA); strapdown sensor assembly; supply support activity; supply support area
SSB	single side band; support services branch; surveillance support branch
SSBN	fleet ballistic missile submarine
SSB-SC	single sideband-suppressed carrier
SSC	small scale contingency; surveillance support center
SSCO	shipper's service control office
SSCRA	Soldiers and Sailors Civil Relief Act
SSE	satellite communications (SATCOM) systems expert; sensitive site exploitation
SSF	software support facility
SSI	standing signal instruction
SSM	surface-to-surface missile
SSMI	special sensor microwave imager
SSMS	single shelter message switch
SSN	attack submarine, nuclear; Social Security number Space surveillance network
SS (number)	sea state (number)
SSO	special security office(r); spot security office
SSP	signals intelligence (SIGINT) support plan
SSPM	single service postal manager
SSPO	strategic systems program office
SSS	Selective Service System; shelter subsystem
SSSC	surface, subsurface search surveillance coordination
SST	space support team; special support team (National Security Agency)
SSTR	stability, security, transition, and reconstruction
ST	short ton; small tug; special tactics; strike team
S/T	short ton
ST&E	security test and evaluation
STA	system tape A
STAB	space tactical awareness brief
STA clk	station clock
STAMMIS	standard Army multi-command management information system
STAMP	standard air munitions package (USAF)
STANAG	standardization agreement (NATO)
STANAVFORLANT	Standing Naval Forces, Atlantic (NATO)
STAR	scheduled theater airlift route; sensitive target approval and review; standard attribute reference; standard terminal arrival route; surface-to-air recovery; system threat assessment report
STARC	state area coordinators
STARS	Standard Accounting and Reporting System
START	Strategic Arms Reduction Treaty
STARTEX	start of exercise
STB	super tropical bleach
STC	secondary traffic channel

STDM	synchronous time division multiplexer
STE	secure telephone equipment
STEL STU III	Standford telecommunications (secure telephone)
STEP	software test and evaluation program; standardized tactical entry point; standard tool for employment planning
STG	seasonal target graphic
STICS	scalable transportable intelligence communications system
STO	special technical operations
STOC	special technical operations coordinator
STOD	special technical operations division
STOL	short takeoff and landing
STOMPS	stand-alone tactical operational message processing system
STON	short ton
STOVL	short takeoff and vertical landing aircraft
STP	security technical procedure
STR	strength
STRAPP	standard tanks, racks and pylons packages (USAF)
STRATOPS	strategic operations division
STREAM	standard tensioned replenishment alongside method
STS	special tactics squadron
STT	small tactical terminal; special tactics team
STU	secure telephone unit
STU-III	secure telephone unit III
STW	strike warfare
STWC	strike warfare commander
STX	start of text
SU	search unit
SUBJ	subject
sub-JIB	subordinate-joint information bureau
SUBOPAUTH	submarine operating authority
sub-PIC	subordinate-press information center
SUBROC	submarine rocket
SUC	surf current
SUIC	service unit identification code
SUMMITS	scenario unrestricted mobility model of intratheater simulation
SUPE	supervisory commands program
SURG	surgeon
SUROBS	surf observation
SURPIC	surface picture
SUW	surface warfare
SUWC	surface warfare commander
S/V	sailboat
SVC	Service
SVIP	secure voice improvement program
SVLTU	service line termination unit
SVR	surface vessel radar
SVS	secure voice system
Sw	switch
SWA	Southwest Asia
SWAT	special weapons and tactics

SWBD	switchboard
SWC	strike warfare commander; swell/wave current
SWI	special weather intelligence
SWO	staff weather officer
SWORD	submarine warfare operations research division
SWSOCC	Southwest Sector Operation Control Center North American Aerospace Defense Command (NORAD)
SWXS	Space Weather Squadron
SYDP	six year defense plan
SYG	Secretary General (UN)
SYNC	synchronization
SYS	system
SYSCON	systems control
SZ	surf zone
T	time (search time available); short ton; trackline pattern
T&DE	test and diagnostic equipment
T&E	test and evaluation
T2	technology transfer
TA	target acquisition; target audience; technical arrangement; theater Army; threat assessment
TAA	tactical assembly area
TAACOM	theater Army area command
TAADS	The Army Authorization Document System
TAAMDCOORD	theater Army air and missile defense coordinator
TAB	tactical air base
TAC	tactical advanced computer; terminal access controller; terminal attack control; terminal attack controller
TAC(A)	tactical air coordinator (airborne)
TACAIR	tactical air
TACAMO	take charge and move out (E-6A/B aircraft)
TACAN	tactical air navigation
TACC	tactical air command center (Marine Corps); tactical air control center (Navy); tanker airlift control center
TAC-D	tactical deception
TACDAR	tactical detection and reporting
TACINTEL	tactical intelligence
TACLAN	tactical local area network
TACLOG	tactical-logistical
TACM	tactical air command manual
TACO	theater allied contracting office
TACON	tactical control
TACOPDAT	tactical operational data
TA/CP	technology assessment/control plan
TACP	tactical air control party
TACRON	tactical air control squadron
T-ACS	auxiliary crane ship
TACS	tactical air control system; theater air control system
TACSAT	tactical satellite
TACSIM	tactical simulation

TACSTANS	tactical standards
TACT	tactical aviation control team
TACTRAGRULANT	Tactical Training Group, Atlantic
TAD	tactical air direction; temporary additional duty (non-unit-related personnel); theater air defense; time available for delivery
TADC	tactical air direction center
TADCS	tactical airborne digital camera system
TADIL	tactical digital information link
TADL	tactical digital information link
TADS	Tactical Air Defense System; target acquisition system and designation sight
TAES	theater aeromedical evacuation system
TAF	tactical air force
TAFDS	tactical airfield fuel dispensing system
TAFIM	technical architecture framework for information management
TAFS	tactical aerodrome forecasts
TAFT	technical assistance field team
TAG	technical assessment group; the adjutant general; Tomahawk land-attack missile aimpoint graphic
T-AGOS	tactical auxiliary general ocean surveillance
TAGS	theater air ground system
T-AH	hospital ship
TAI	International Atomic Time; target area of interest; total active inventory
TAIS	transportation automated information systems
TAK	cargo ship
T-AKR	fast logistics ship
TALCE	tanker airlift control element
TALD	tactical air-launched decoy
TALON	Threat and Local Observation Notice
TAMCA	theater Army movement control agency
TAMCO	theater Army movement control center
TAMD	theater air and missile defense
TAMMC	theater army material management command
TAMMIS	theater Army medical management information system
TAMS	transportation analysis, modeling, and simulation
tanalt	tangent altitude
TAO	tactical actions officer
TAOC	tactical air operations center (USMC)
TAP	troopship
TAR	tactical air request; Training and Administration of the Reserve
TARBS	transportable amplitude modulation and frequency modulation radio broadcast system
TARBUL	target bulletin
TARE	tactical record evaluation
TAREX	target exploitation; target plans and operations
TARS	tethered aerostat radar system
TARWI	target weather and intelligence
TAS	tactical atmospheric summary; true air speed
T-ASA	Television Audio Support Agency

TASCID	tactical Automatic Digital Network (AUTODIN) satellite compensation interface device
TASCO	tactical automatic switch control officer
TASIP	tailored analytic intelligence support to individual electronic warfare and command and control warfare projects
TASKORD	tasking order
TASMO	tactical air support for maritime operations
TASOSC	theater Army special operations support command
TASS	tactical automated security system; tactical automated switch system
TAT	tactical analysis team technical assistance team
TATC	tactical air traffic control
T-AVB	aviation logistics support ship
TAW	tactical airlift wing
TBD	to be determined
TBM	tactical ballistic missile; theater ballistic missile
TBMCS	theater battle management core system
TBMD	theater ballistic missile defense
TBP	to be published
TBSL	to be supplied later
TBTC	transportable blood transshipment center
TC	tidal current; transmit clock and/or telemetry combiner; training circular; Transportation Corps (Army)
TCA	terminal control area; time of closest approach; traditional combatant commander activity
TC-ACCIS	Transportation Coordinator's Automated Command and Control Information System
TC-AIMS	Transportation Coordinator's Automated Information for Movement System
TC-AIMS II	Transportation Coordinator's Automated Information for Movement System II
TCAM	theater Army medical management information system (TAMMIS) customer assistance module
TCC	transmission control code; transportation component command
TCCF	tactical communications control facility
TCEM	theater contingency engineering management
TCF	tactical combat force; technical control facility
TCM	theater construction manager
TCMD	transportation control and movement document
TCN	third country national; transportation control number
TCS	theater communications system
TCSEC	trusted computer system evaluation criteria
TD	temporary duty; theater distribution; timing distributor; total drift; transmit data
TDA	Table of Distribution and Allowance
TDAD	Table of Distribution and Allowance (TDA) designation
T-day	effective day coincident with Presidential declaration of a National Emergency and authorization of partial mobilization
TDBM	technical database management
TDBSS	Theater Defense Blood Standard System
TDD	target desired ground zero (DGZ) designator; time definite delivery

TDF	tactical digital facsimile
TDI	target data inventory
TDIC	time division interface controller
TDIG	time division interface group
TDIM	time division interface module
TDL	tactical data link
TDM	time division multiplexed
TDMA	time division multiple access
TDMC	theater distribution management cell
TDMF	time division matrix function
TDMM	time division memory module
TDMX	time division matrix
TDN	target development nomination
TDP	theater distribution plan
TDR	transportation discrepancy report
TDSG	time division switching group
TDSGM	time division switching group modified
TDT	theater display terminal
TDY	temporary duty
TE	transaction editor
TEA	Transportation Engineering Agency
Tech	technical
TECHCON	technical control
TECHDOC	technical documentation
TECHELINT	technical electronic intelligence
TECHEVAL	technical evaluation
TECHINT	technical intelligence
TECHOPDAT	technical operational data
TECS II	Treasury Enforcement Communications System
TED	trunk encryption device
TEK	TeleEngineering Kit
TEL	transporter-erector-launcher (missile platform)
TELEX	teletype
TELINT	telemetry intelligence
TELNET	telecommunication network
TEMPER	tent extendible modular personnel
TENCAP	tactical exploitation of national capabilities program
TEOB	tactical electronic order of battle
TEP	test and evaluation plan; theater engagement plan
TERCOM	terrain contour matching
TERF	terrain flight
TERPES	tactical electronic reconnaissance processing and evaluation system
TERPROM	terrain profile matching
TERS	tactical event reporting system
TES	tactical event system; theater event system
TESS	Tactical Environmental Support System
TEU	technical escort unit; twenty-foot equivalent unit
TEWLS	Theater Enterprise Wide Logistics System
TF	task force
TFA	toxic free area

TFADS	Table Formatted Aeronautic Data Set
TFCICA	task force counterintelligence coordinating authority
TFE	tactical field exchange; transportation feasibility estimator
TFLIR	targeting forward-looking infrared
TFMS-M	Transportation Financial Management System-Military
TFR	temporary flight restriction
TFS	tactical fighter squadron; Tactical Forecast System
TG	task group
TGC	trunk group cluster
TGEN	table generate
TGM	trunk group multiplexer
TGMOW	transmission group module and/or order wire
TGO	terminal guidance operations
TGT	target
TGTINFOREP	target information report
TGU	trunk compatibility unit
TI	threat identification; training instructor
TIAP	theater intelligence architecture program
TIARA	tactical intelligence and related activities
TIB	toxic industrial biological
TIBS	tactical information broadcast service
TIC	target information center; toxic industrial chemical
TIDP	technical interface design plan
TIDS	tactical imagery dissemination system
TIFF	tagged image file format
TII	total inactive inventory
TIM	theater information management; toxic industrial material
TIO	target intelligence officer
TIP	target intelligence package
TIPG	telephone interface planning guide
TIPI	tactical information processing and interpretation system; tactical information processing interpretation
TIPS	tactical optical surveillance system (TOSS) imagery processing system
TIR	toxic industrial radiological
TIROS	television infrared observation satellite
TIS	technical interface specification; thermal imaging system
TISG	technical interoperability standards group
TISS	thermal imaging sensor system
TJAG	the judge advocate general
T-JMC	theater-joint movement center
T-JTB	Theater-Joint Transportation Board
TJTN	theater joint tactical network
TL	team leader
TLAM	Tomahawk land attack missile
TLAMM	theater lead agent for medical materiel
TLAM/N	Tomahawk land attack missile/nuclear
TLC	traffic load control
TLCF	teleconference
TLE	target location error
TLM	topographic line map

TLP	transmission level point
TLR	trailer
TLX	teletype
TM	tactical missile; target materials; team member; technical manual; theater missile; TROPO modem
TMAO	theater mortuary affairs officer
TMD	tactical munitions dispenser; theater missile defense
TMEP	theater mortuary evacuation point
TMG	timing
TMIP	theater medical information program
TMIS	theater medical information system
TML	terminal
TMLMC	theater medical logistic management center
TMMMC	theater medical materiel management center
TMN	trackline multiunit non-return
TMO	traffic management office; transportation management office
TMP	target materials program; Telecommunications management program; theater manpower forces
TMR	trackline multiunit return
T/M/S	type, model, and/or series (also as TMS)
TNAPS	tactical network analysis and planning system
TNAPS+	tactical network analysis and planning system plus
TNC	theater network operations (NETOPS) center
TNCC	theater network operations (NETOPS) control center
TNCO	transnational criminal organization
T-net	training net
TNF	theater nuclear force
TNL	target nomination list
T/O	table of organization
TO	technical order; theater of operations
TO&E	table of organization and equipment
TOA	table of allowance
TOAI	total overall aircraft inventory
TOC	tactical operations center; tanker (airlift) control center (TALCE) operations center
TOCU	tropospheric scatter (TROPO) orderwire control unit
TOD	time of day
TOE	table of organization and equipment
TOF	time of flight
TOFC	trailer on flatcar
TOH	top of hill
TOI	track of interest
TOPINT	technical operational intelligence
TOR	term of reference; time of receipt
TOS	time on station
TOSS	tactical optical surveillance system
TOT	time on target
TOW	tube launched, optically tracked, wire guided
TP	technical publication; turn point
TPB	tactical psychological operations battalion

TPC	tactical pilotage chart; two person control
TPC/PC	tactical pilotage chart and/or pilotage chart
TPED	tasking, processing, exploitation, and dissemination
TPERS	type personnel element
TPFDD	time-phased force and deployment data
TPFDL	time-phased force and deployment list
TPL	technical publications list; telephone private line
TPME	task, purpose, method, and effects
TPMRC	theater patient movement requirements center
TPO	task performance observation
TPRC	theater planning response cell
TPT	tactical petroleum terminal
TPTRL	time-phased transportation requirements list
TPU	tank pump unit
TRA	technical review authority
TRAC2ES	transportation command regulating and command and control evacuation system
TRACON	terminal radar approach control facility
TRADOC	United States Army Training and Doctrine Command
TRAM	target recognition attack multisensor
TRANSEC	transmission security
TRAP	tactical recovery of aircraft and personnel (Marine Corps); tactical related applications; tanks, racks, adapters, and pylons; terrorism research and analysis program
TRC	tactical radio communication; transmission release code
TRCC	tactical record communications center
TRE	tactical receive equipment
TREAS	Department of the Treasury
TREE	transient radiation effects on electronics
TRICON	triple container
TRI-TAC	Tri-Service Tactical Communications Program
TRK	truck; trunk
TRNG	training
TRO	training and readiness oversight
TROPO	troposphere; tropospheric scatter
TRS	tactical reconnaissance squadron
TS	terminal service; top secret
TSA	target system analysis; theater storage area; Transportation Security Administration; travel security advisory
TSB	technical support branch; trunk signaling buffer
TSBn	transportation support battalion (USMC)
TSC	theater security cooperation; theater support command
TSCIF	tactical sensitive compartmented information facility
TSCM	technical surveillance countermeasures
TSCO	target selection confusion of the operator; top secret control officer
TSCP	theater security cooperation plan
TSCR	time sensitive collection requirement
TSE	tactical support element
TSEC	transmission security
TSG	targeting support group; test signal generator

TSGCE	tri-Service group on communications and electronics
TSGCEE	tri-Service group on communications and electronic equipment (NATO)
TSM	trunk signaling message
TSN	trackline single-unit non-return; track supervision net
TSO	technical standard order; telecommunications service order
TSOC	theater special operations command
TSP	telecommunications service priority
TSR	telecommunications service request; theater source registry; theater support representative; trackline single-unit return
TSS	tactical shelter system; target sensing system; timesharing system; time signal set; traffic service station
TSSP	tactical satellite signal processor
TSSR	tropospheric scatter (TROPO)-satellite support radio
TST	tactical support team; theater support team; time-sensitive target
TSWA	temporary secure working area
TT	terminal transfer
TT&C	telemetry, tracking, and commanding
TTB	transportation terminal battalion
TTD	tactical terrain data; technical task directive
TTG	thermally tempered glass
TTL	transistor-transistor logic
TTM	threat training manual; training target material
TTP	tactics, techniques, and procedures; trailer transfer point
TTR	tactical training range
TTT	time to target
TTU	transportation terminal unit
TTY	teletype
TUBA	transition unit box assembly
TUCHA	type unit characteristics file
TUCHAREP	type unit characteristics report
TUDET	type unit equipment detail file
TV	television
TVA	Tennessee Valley Authority
TW/AA	tactical warning and attack assessment
TWC	Office for Counterterrorism Analysis (DIA); total water current
TWCF	Transportation Working Capital Fund
TWCM	theater wartime construction manager
TWD	transnational warfare counterdrug analysis
TWDS	tactical water distribution system
TWI	Office for Information Warfare Support (DIA)
TWPL	teletypewriter private line
TWX	teletypewriter exchange
TX	transmitter; transmit
TYCOM	type commander
U	wind speed
UA	unmanned aircraft
UAOBS	upper air observation
UAR	unconventional assisted recovery

UARCC	unconventional assisted recovery coordination cell
UARM	unconventional assisted recovery mechanism
UART	unconventional assisted recovery team
UAS	unmanned aircraft system
UAV	unmanned aerial vehicle
U/C	unit cost; upconverter
UCFF	Unit Type Code Consumption Factors File
UCMJ	Uniform Code of Military Justice
UCP	Unified Command Plan
UCT	underwater construction team
UDAC	unauthorized disclosure analysis center
UDC	unit descriptor code
UDESC	unit description
UDL	unit designation list
UDP	unit deployment program
UDT	underwater demolition team
UE	unit equipment
UFC	Unified Facilities Criteria
UFO	ultrahigh frequency follow-on
UFR	unfunded requirement
UGM-84A	Harpoon
UGM-96A	Trident I
UGV	Unmanned Ground Vehicle
UHF	ultrahigh frequency
UHV	Upper Huallaga Valley
UIC	unit identification code
UICIO	unit identification code information officer
UIRV	unique interswitch rekeying variable
UIS	unit identification system
UJTL	Universal Joint Task List
UK	United Kingdom
UK(I)	United Kingdom and Ireland
ULC	unit level code
ULF	ultra low frequency
ULLS	unit level logistics system
ULN	unit line number
UMCC	unit movement control center
UMCM	underwater mine countermeasures
UMD	unit manning document; unit movement data
UMIB	urgent marine information broadcast
UMMIPS	uniform material movement and issue priority system
UMPR	unit manpower personnel record
UMT	unit ministry team
UN	United Nations
UNAMIR	United Nations Assistance Mission in Rwanda
UNC	United Nations Command
UNCTAD	United Nations Conference on Trade and Development
UND	urgency of need designator
UNDHA	United Nations Department of Humanitarian Affairs
UN-DMT	United Nations disaster management team

UNDP	United Nations development programme
UNDPKO	United Nations Department for Peacekeeping Operations
UNEF	United Nations emergency force
UNEP	United Nations environment program
UNESCO	United Nations Educational, Scientific, and Cultural Organization
UNHCHR	United Nations High Commissioner for Human Rights
UNHCR	United Nations Office of the High Commissioner for Refugees
UNICEF	United Nations Children's Fund
UNIFIL	United Nations Interim Force in Lebanon
UNIL	unclassified national information library
UNITAF	unified task force
UNITAR	United Nations Institute for Training and Research
UNITREP	unit status and identity report
UNLOC	United Nations logistic course
UNMIH	United Nations Mission in Haiti
UNMILPOC	United Nations military police course
UNMOC	United Nations military observers course
UNMOVCC	United Nations movement control course
UNO	unit number
UNOCHA	United Nations Office for the Coordination of Humanitarian Affairs
UNODC	United Nations Office on Drugs and Crime
UNODIR	unless otherwise directed
UNOSOM	United Nations Operations in Somalia
UNPA	United Nations Participation Act
UNPROFOR	United Nations protection force
UNREP	underway replenishment
UNREP CONSOL	underway replenishment consolidation
UNRWA	United Nations Relief and Works Agency for Palestine Refugees in the Near East
UNSC	United Nations Security Council
UNSCR	United Nations Security Council resolution
UNSG	United Nations Secretary-General
UNSOC	United Nations staff officers course
UNTAC	United Nations Transition Authority in Cambodia
UNTSO	United Nations Truce and Supervision Organization
UNV	United Nations volunteer
UP&TT	unit personnel and tonnage table
UPU	Universal Postal Union
URDB	user requirements database
USA	United States Army
USAB	United States Army barracks
USACCSA	United States Army Command and Control Support Agency
USACE	United States Army Corps of Engineers
USACFSC	United States Army Community and Family Support Center
USACHPPM	US Army Center for Health Promotion and Preventive Medicine
USACIDC	United States Army Criminal Investigation Command
USAF	United States Air Force
USAFE	United States Air Forces in Europe
USAFEP	United States Air Force, Europe pamphlet
USAFLANT	United States Air Force, Atlantic Command

USAFR	United States Air Force Reserve
USAFSOC	United States Air Force, Special Operations Command
USAFSOF	United States Air Force, Special Operations Forces
USAFSOS	USAF Special Operations School
USAID	United States Agency for International Development
USAITAC	United States Army Intelligence Threat Analysis Center
USAJFKSWC	United States Army John F. Kennedy Special Warfare Center
USAMC	United States Army Materiel Command
USAMMA	United States Army Medical Materiel Agency
USAMPS	United States Army Military Police School
USAMRICD	US Army Medical Research Institute for Chemical Defense
USAMRIID	US Army Medical Research Institute of Infectious Diseases
USAMRMC	US Army Medical Research and Materiel Command
USAO	United States Attorney Office
USAR	United States Army Reserve
USARCENT	United States Army, Central Command
USAREUR	United States Army, European Command
USARIEM	United States Army Research Institute of Environmental Medicine
USARJ	United States Army, Japan
USARNORTH	US Army Forces North
USARPAC	United States Army, Pacific Command
USARSO	United States Army, Southern Command
USASOC	US Army Special Operations Command
USB	upper side band Universal Serial Bus
USBP	United States Border Patrol
USC	United States Code; universal service contract
USCENTAF	United States Central Command Air Forces
USCENTCOM	United States Central Command
USCG	United States Coast Guard
USCGR	United States Coast Guard Reserve
USCIS	US Citizenship and Immigration Services
USCS	United States Cryptologic System; United States Customs Service
USDA	United States Department of Agriculture
USD(A&T)	Under Secretary of Defense for Acquisition and Technology
USDAO	United States defense attaché office
USD(AT&L)	Under Secretary of Defense for Acquisition, Technology, and Logistics
USD(C)	Under Secretary of Defense (Comptroller)
USDELMC	United States Delegation to the NATO Military Committee
USD(I)	Under Secretary of Defense (Intelligence)
USD(P)	Under Secretary of Defense for Policy
USD(P&R)	Under Secretary of Defense (Personnel & Readiness)
USDR	United States defense representative
USD(R&E)	Under Secretary of Defense for Research and Engineering
USELEMCMOC	United States Element Cheyenne Mountain Operations Center
USELEMNORAD	United States Element, North American Aerospace Defense Command
USERID	user identification
USEUCOM	United States European Command
USFJ	United States Forces, Japan
USFK	United States Forces, Korea

USFORAZORES	United States Forces, Azores
USFS	United States Forest Service
USFWS	United States Fish and Wildlife Service
USG	United States Government
USGS	United States Geological Survey
USIA	United States Information Agency
USIC	United States interdiction coordinator
USIS	United States Information Service
USJFCOM	United States Joint Forces Command
USLANTFLT	United States Atlantic Fleet
USLO	United States liaison officer
USMARFORCENT	United States Marine Component, Central Command
USMARFORLANT	United States Marine Component, Atlantic Command
USMARFORPAC	United States Marine Component, Pacific Command
USMARFORSOUTH	United States Marine Component, Southern Command
USMC	United States Marine Corps
USMCEB	United States Military Communications-Electronics Board
USMCR	United States Marine Corps Reserve
USMER	United States merchant ship vessel locator reporting system
USMILGP	United States military group
USMILREP	United States military representative
USMOG-W	United States Military Observer Group - Washington
USMS	United States Marshals Service
USMTF	United States message text format
USMTM	United States military training mission
USN	United States Navy
USNAVCENT	United States Naval Forces, Central Command
USNAVEUR	United States Naval Forces, Europe
USNAVSO	US Naval Forces Southern Command
USNCB	United States National Central Bureau (INTERPOL)
USNMR	United States National Military representative
USNMTG	United States North Atlantic Treaty Organization (NATO) Military Terminology Group
USNORTHCOM	United States Northern Command
USNR	United States Navy Reserve
USNS	United States Naval Ship
USPACAF	United States Air Forces, Pacific Command
USPACFLT	United States Pacific Fleet
USPACOM	United States Pacific Command
USPHS	United States Public Health Service
USPS	United States Postal Service
USREPMC	United States representative to the military committee (NATO)
USSOCOM	United States Special Operations Command
USSOUTHAF	United States Air Force, Southern Command
USSOUTHCOM	United States Southern Command
USSPACECOM	United States Space Command
USSS	United States Secret Service (TREAS); United States Signals Intelligence (SIGINT) System
USSTRATCOM	United States Strategic Command
USTRANSCOM	United States Transportation Command

USUN	United States Mission to the United Nations
USW	undersea warfare
USW/USWC	undersea warfare and/or undersea warfare commander
USYG	Under Secretary General
UT1	unit trainer; Universal Time
UTC	Coordinated Universal Time; unit type code
UTM	universal transverse mercator
UTO	unit table of organization
UTR	underwater tracking range
UUV	unmanned underwater vehicle
UVEPROM	ultraviolet erasable programmable read-only memory
UW	unconventional warfare
UWOA	unconventional warfare operating area
UXO	unexploded explosive ordnance; unexploded ordnance
V	search and rescue unit ground speed; sector pattern; volt v velocity of target drift
VA	Veterans Administration; victim advocate; vulnerability assessment
VAAP	vulnerability assessment and assistance program
VAC	volts, alternating current
VARVAL	vessel arrival data, list of vessels available to marine safety offices and captains of the port
VAT B	(weather) visibility (in miles), amount (of clouds, in eighths), (height of cloud) top (in thousands of feet), (height of cloud) base (in thousands of feet)
VBIED	vehicle-borne improvised explosive device
VBS	visit, board, search
VBSS	visit, board, search, and seizure
VCC	voice communications circuit
VCG	virtual coordination group
VCJCS	Vice Chairman of the Joint Chiefs of Staff
VCNOG	Vice Chairman, Nuclear Operations Group
VCO	voltage controlled oscillator
VCOPG	Vice Chairman, Operations Planners Group
VCR	violent crime report
VCXO	voltage controlled crystal oscillator; voltage controlled oscillator
VDC	volts, direct current
VDJS	Vice Director, Joint Staff
VDR	voice digitization rate
VDS	video subsystem
VDSD	visual distress signaling device
VDU	visual display unit
VDUC	visual display unit controller
VEH	vehicle; vehicular cargo
VERTREP	vertical replenishment
VF	voice frequency
VFR	visual flight rules
VFS	validating flight surgeon
VFTG	voice frequency telegraph
VHF	very high frequency

VI	visual information
VICE	advice
VID	visual identification information display
VIDOC	visual information documentation
VINSON	encrypted ultrahigh frequency communications system
VIP	very important person; visual information processor
VIRS	verbally initiated release system
VIS	visual imaging system
VISA	Voluntary Intermodal Sealift Agreement
VISOBS	visual observer
VIXS	video information exchange system
VLA	vertical line array; visual landing aid
VLF	very low frequency
VLR	very-long-range aircraft
VLZ	vertical landing zone
VMap	vector map
VMAQ	Marine tactical electronic warfare squadron
VMC	visual meteorological conditions
VMF	variable message format
VMGR	Marine aerial refueler and transport squadron
VMI	vendor managed inventory
VNTK	target vulnerability indicator designating degree of hardness; susceptibility of blast; and K-factor VO validation office
VOCODER	voice encoder
VOCU	voice orderwire control unit
VOD	vertical onboard delivery
VOL	volunteer
Vol	volume
VOLS	vertical optical landing system
VOR	very high frequency omnidirectional range station
VORTAC	very high frequency omnidirectional range station and/or tactical air navigation
VOX	voice actuation (keying)
VP	video processor
VPB	version planning board
VPD	version planning document
VPV	virtual prime vendor
VS	sector single-unit
VS&PT	vehicle summary and priority table
VSAT	very small aperture terminal
VSG	virtual support group
VSII	very seriously ill or injured
VSP	voice selection panel
VSR	sector single-unit radar
V/STOL	vertical and/or short takeoff and landing aircraft
VSW	very shallow water
VTA	voluntary tanker agreement
VTC	video teleconferencing
VTOL	vertical takeoff and landing
VTOL-UAV	vertical takeoff and landing unmanned aerial vehicle

VTS	vessel traffic service
VTT	video teletraining
VU	volume unit
VV&A	verification, validation, and accreditation
VV&C	verification, validation, and certification
VX	nerve agent (O-Ethyl S-Diisopropylaminomethyl Methylphosphonothiolate)
W	sweep width
w	search subarea width
WAAR	Wartime Aircraft Activity Report
WACBE	World Area Code Basic Encyclopedia
WADS	Western Air Defense Sector
WAGB	icebreaker (USCG)
WAN	wide-area network
WAR	Weekly Activity Report
WARM	wartime reserve mode
WARMAPS	wartime manpower planning system
WARNORD	warning order
WARP	web-based access and retrieval portal
WAS	wide area surveillance
WASP	war air service program
WATCHCON	watch condition
WB	wideband
WC	wind current
WCA	water clearance authority
WCCS	Wing Command and Control System
WCDO	War Consumables Distribution Objective
WCO	World Customs Organization
WCS	weapons control status
W-day	declared by the President, W-day is associated with an adversary decision to prepare for war
WDT	warning and display terminal
WEAX	weather facsimile
WES	weapon engagement status
WETM	weather team
WEU	Western European Union
WEZ	weapon engagement zone
WFE	warfighting environment
WFP	World Food Programme (UN)
WG	working group
WGS	World Geodetic System
WGS-84	World Geodetic System 1984
WH	wounded due to hostilities
WHEC	high-endurance cutter (USCG)
WHNRS	wartime host-nation religious support
WHNS	wartime host-nation support
WHNSIMS	Wartime Host Nation Support Information Management System
WHO	World Health Organization (UN)

WIA	wounded in action
WISDIM	Warfighting and Intelligence Systems Dictionary for Information Management
WISP	Wartime Information Security Program
WIT	weapons intelligence team
WLG	Washington Liaison Group
WMD	weapons of mass destruction
WMD/CM	weapons of mass destruction consequence management
WMD-CST	weapons of mass destruction-civil support team
WMEC	Coast Guard medium-endurance cutter
WMO	World Meteorological Organization
WMP	Air Force War and Mobilization Plan; War and Mobilization Plan
WOC	wing operations center (USAF)
WOD	word-of-day
WORM	write once read many
WOT	war on terrorism
WP	white phosphorous; Working Party (NATO)
WPA	water jet propulsion assembly
WPAL	wartime personnel allowance list
WPARR	War Plans Additive Requirements Roster
WPB	Coast Guard patrol boat
WPC	Washington Planning Center
WPM	words per minute
WPN	weapon
WPR	War Powers Resolution
WPS	Worldwide Port System
WR	war reserve; weapon radius
WRAIR	Walter Reed Army Institute of Research
WRC	World Radiocommunication Conference
WRL	weapons release line
WRM	war reserve materiel
WRMS	war reserve materiel stock
WRR	weapons response range (as well as wpns release rg)
WRS	war reserve stock
WRSK	war readiness spares kit; war reserve spares kit
WSE	weapon support equipment
WSES	surface effect ship (USCG)
WSESRB	Weapon System Explosive Safety Review Board
WSR	weapon system reliability
WSV	weapon system video
WT	gross weight; warping tug; weight
WTCA	water terminal clearance authority
WTCT	weapons of mass destruction technical collection team
WTLO	water terminal logistic office
Wu	uncorrected sweep width
WVRD	World Vision Relief and Development, Inc.
WWABNCP	worldwide airborne command post
WWII	World War II
WWSVCS	Worldwide Secure Voice Conferencing System
WWX	worldwide express

WX	weather
X	initial position error
XCVR	transceiver
XO	executive officer
XSB	barrier single unit
Y	search and rescue unit (SRU) error
YR	year
Z	zulu
z	effort
ZF	zone of fire
Zt	total available effort
ZULU	time zone indicator for Universal Time

Index

A

B

C

D

E

I

J

K

L

M

N

O

P

R

S

T

V

W

Y

Z

Manufacturing & Industrial Engineering

In light of increasing economic and international threats, military operations must be examined with a critical eye in terms of process design, management, improvement, and control. Although the Pentagon and militaries around the world have utilized industrial engineering (IE) concepts to achieve this goal for decades, there has been no single resource to bring together IE applications with a focus on improving military operations. Until now…

The **Handbook of Military Industrial Engineering** is the first compilation of the fundamental tools, principles, and modeling techniques of industrial engineering with specific and direct application to military systems. Globally respected IE experts provide proven strategies that can help any military organization effectively create, adapt, utilize, and deploy resources, tools, and technology.

Topics covered include:

- Supply chain management and decision making
- Lean enterprise concepts for military operations
- Modeling and optimization
- Economic planning for military systems
- Contingency planning and logistics
- Human factors and ergonomics
- Information management and control

Civilian engineers working on systems analysis, project management, process design, and operations research will also find inspiration and useful ideas on how to effectively apply the concepts covered for non-military uses.

On the battlefield and in business, victory goes to those who utilize their resources most effectively, especially in times of operational crisis. The **Handbook of Military Industrial Engineering** is a complete reference that will serve as an invaluable resource for those looking to make the operational improvements needed to accomplish the mission at hand.

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

66285
ISBN: 978-1-4200-6628-9
90000
9 781420 066289