MANAGING BUSINESS PROCESS FLOWS

PRINCIPLES OF OPERATIONS MANAGEMENT

THIRD EDITION

Ravi Anupindi | Sunil Chopra | Sudhakar D. Deshmukh
Jan A. Van Mieghem | Eitan Zemel

Managing Business Process Flows

This page intentionally left blank

Third Edition

Managing Business Process Flows

PRINCIPLES OF OPERATIONS MANAGEMENT

Ravi Anupindi
University of Michigan

Sunil Chopra
Northwestern University

Sudhakar D. Deshmukh
Northwestern University

Jan A. Van Mieghem
Northwestern University

Eitan Zemel
New York University

Prentice Hall
Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editorial Director: Sally Yagan
Editor in Chief: Eric Svendsen
Senior Acquisitions Editor: Chuck Synovec
Senior Project Manager: Mary Kate Murray
Editorial Assistant: Ashlee Bradbury
Director of Marketing: Patrice Lumumba Jones
Executive Marketing Manager: Anne Fahlgren
Production Project Manager: Clara Bartunek
Creative Art Director: Jayne Conte
Cover Designer: Suzanne Duda
Cover Art: Fotolia
Manager, Rights and Permissions: Hessa Albader
Media Project Manager: John Cassar
Media Product Manager: Sarah Peterson
Full-Service Project Management: Mohinder Singh/Aptara®, Inc.
Printer/Binder: Edwards Brothers Incorporated
Cover Printer: Lehigh-Phoenix Color/Hagerstown
Text Font: Palatino

Credits and acknowledgments borrowed from other sources and reproduced, with permission, in this textbook appear on appropriate page within text.

Library of Congress Cataloging-in-Publication Data

Managing business process flows : principles of operations management / Ravi Anupindi . . . [et al.].—3rd ed.
p. cm.
Includes index.
ISBN-13: 978-0-13-603637-1
ISBN-10: 0-13-603637-6
1. Production management. 2. Process control. I. Anupindi, Ravi.
TS155.M33217 2012
658.5—dc23

2011023366

10 9 8 7 6 5 4 3 2 1

PEARSON

ISBN 10: 0-13-603637-6
ISBN 13: 978-0-13-603637-1

BRIEF CONTENTS

This page intentionally left blank

CONTENTS

This page intentionally left blank

PREFACE

In this book, we present a novel approach to studying the core concepts in operations, which is one of the three major functional fields in business management, along with finance and marketing. We view the task, and the raison d'être, of operations management as structuring (designing), managing, and improving organizational *processes* and use the process view as the unifying paradigm to study operations. We address manufacturing as well as service operations in make-to-stock as well as make-to-order environments.

We employ a structured data-driven approach to discuss the core operations management concepts in three steps:

- Model and understand a business process and its flows.
- Study causal relationships between the process structure and operational and financial performance metrics.
- Formulate implications for managerial actions by filtering out managerial "levers" (process drivers) and their impact on operational and financial measures of process performance.

NEW TO THIS EDITION

The first edition of this book was published in 1999 and reflected our experiences from teaching the core Operations Management course at the Kellogg School of Management of Northwestern University. The second edition, published in 2006, improved exposition and clarified the link between theory and practice. While this third edition retains the general process-view paradigm, we have striven to sharpen the development of the ideas in each chapter, illustrate with contemporary examples from practice, and eliminated some content to make room for some new content, such as:

- Opening vignettes and real-life examples of how the theory can be applied in practice have been made current. In addition, exposition of material in the chapters has been further improved with technical derivations details and other tangential ideas relegated to chapter appendices.
- Chapter 4 has been completely revised, with an emphasis on measurement, analysis of critical path, and management approaches to leadtime improvements. Technical analysis has been shifted to appendices.
- Chapter 5 has been substantially revised with emphasis on effective capacity and bottleneck management, on the effects of product mix on capacity, and on reduction of capacity waste.
- Chapter 6 now includes discussion of quantity discount policies. Discussions of periodic review policies have been added to Chapters 6 and 7.
- Chapter 8 has undergone a complete revision and reorganization to improve flow of concepts; we have also added some discussion on priority processing.
- Chapter 9 has more details on control charts, includes fraction defective chart, recent applications, discussion of integrated design, and total quality management.
- Answers to selected exercises from Chapters 3 to 9 appear at the end of the book.
- The end-of-chapter and end-of-book features have been updated.

Finally, we have removed iGrafx simulation (both the software and the associated sample models) from this edition.

OVERVIEW

Our objective is to show how managers can design and manage process structure and process drivers to improve the performance of any business process. The book consists of four parts.

In Part I, "Process Management and Strategy," we introduce the basic concepts of business processes and management strategy. Processes are the core technologies of any organization to design, produce and deliver products and services that satisfy external and internal customer needs. Processes involve transforming inputs into outputs by means of capital and labor resources that carry out a set of interrelated activities. The existence of trade-offs in process competencies implies that world-class operations must align their competencies with the desired product attributes and overall competitive priorities as formulated by the competitive strategy.

In Part II, "Process Flow Metrics," we examine key process measures, their interrelationships, and managerial levers for controlling them. In particular, process flow time, flow rate or throughput, and inventory are three fundamental operational measures that affect the financial measures of process performance. Flow time can be improved by restructuring and shortening the time-critical path of activities; throughput can be improved by increasing the bottleneck capacity, and inventory can be decreased by reducing the batch sizes, streamlining the process, or reducing variability. Yet, throughout this part, the focus is on the average values, ignoring for now the impact of variability in process performance.

In Part III, "Process Flow Variability," we study the effect of variability in flows and processing on the process performance and the managerial levers to plan for and control it. Safety inventory is used to maintain the availability of inputs and outputs in spite of variability in inflows and demands in the make-to-stock environment. Safety capacity is used to minimize waiting times due to variability in inflows and processing times in the make-to-order environment. Safety time is used to provide a reliable estimate of the response time to serve a customer. Finally, feedback control is used to monitor and respond to variability in process performance dynamically over time.

In Part IV, "Process Integration," we conclude with principles of synchronization of flows of materials and information through a network of processes most economically. The ideal is to eliminate waste in the form of excess costs, defects, delays, and inventories. Instead of responding to the economies of scale and variability in flows, the long-term approach is to eliminate the need for such responses by making processes lean, flexible, and predictable. It requires continual exposure and elimination of sources of inefficiency, rigidity, and variability and use of information technology to integrate various subprocesses. The goal is to design and control the process for continuous flows without waits, inventories, and defects. We close with the different philosophies of process improvement toward achieving this goal.

In Appendix I, we give a summary of the "levers" to manage business processes. We hope that this checklist will be useful to the practitioner. We assume that our readers have knowledge of some basic concepts in probability and statistics; for completeness, we summarize these as background material in Appendix II.

INSTRUCTOR RESOURCES

- ***Instructor Resource Center:*** The Instructor Resource Center contains the electronic files for the test bank, PowerPoint slides, and the Solutions Manual. (www.pearsonhighered.com/anupindi).

- ***Register, Redeem, Login:*** At www.pearsonhighered.com/irc, instructors can access a variety of print, media, and presentation resources that are available with this text in downloadable, digital format. For most texts, resources are also available for course management platforms such as Blackboard, WebCT, and Course Compass.
- ***Need help?*** Our dedicated technical support team is ready to assist instructors with questions about the media supplements that accompany this text. Visit http://247.pearsoned.com/ for answers to frequently asked questions and toll-free user support phone numbers. The supplements are available to adopting instructors. Detailed descriptions are provided on the Instructor Resource Center.

Instructor's Solutions Manual

The Instructor's Solutions Manual, updated by the authors, is available to adopters as a download from the Instructor Resource Center.

Test Item File

The test item file, updated by the authors, is available to adopters as a downloaded from the Instructor Resource Center.

PowerPoint Presentations

The PowerPoint presentations, updated by the authors, are available to adopters as a downloaded from the Instructor Resource Center.

ACKNOWLEDGMENTS

We gratefully acknowledge the feedback from our full-time, part-time, and executive management students at our respective institutions and numerous adopters of the textbook at other institutions. Our colleagues Krishnan Anand (now at David Eccles School of Business, University of Utah), Sarang Deo, Martin (Marty) Lariviere, Andy King (now at Dartmouth College), and Matt Tuite (now retired) have, over time, given us many suggestions for improvement. In particular, Anand suggested the original Loan Application Flow example in Chapter 3, while Marty offered us several new exercises. (Instructors know that good problem sets are golden.) Andy pointed out the need to explicitly account for setup times in determining flow rate more accurately. In addition, we also benefited from the suggestions by several colleagues at other universities. We are particularly indebted to Larry Robinson at Cornell University, George Monahan at the University of Illinois at Urbana–Champaign, Kevin Gue and Ken Doerr of the Naval Postgraduate School at Monterey, and Marty Puterman at the University of British Columbia.

The manuscript has benefited significantly from extensive and meticulous reviews from Amy Whitaker, developmental editor at Pearson Prentice Hall. We are thankful to her for suggesting, among other things, the idea of a glossary of terms and helping us prepare this list. Several people from the staff at Pearson Prentice Hall have really worked hard in patiently coordinating the entire project. In particular, we are thankful to Mary Kate Murray, Senior Project Manager; Chuck Synovec, Senior Acquisition Editor; Anne Fahlgren, Executive Marketing Manager; Clara Bartunek, Production Project Manager. We also thank Mohinder Singh of Aptara Incorporation for his assistance with the production of the book.

Finally, all of us have been influenced in various ways by the way we were taught operations at our respective alma maters. Parts of the book reflect what each of us imbibed from the various classes we took. So we thank our mentors and other faculty at Carnegie Mellon University, Stanford University, the State University of New York at Stony Brook, and the University of California at Berkeley. Last, but not least, we would like to thank our families for their support during this effort.

Ravi Anupindi
Stephen M. Ross School of Business
University of Michigan, Ann Arbor

Sunil Chopra,
Sudhakar D. Deshmukh,
and Jan A. Van Mieghem
J.L. Kellogg School of Management
Northwestern University

Eitan Zemel
Leonard N. Stern School of Business
New York University

PART I

Process Management and Strategy

Chapter 1

Products, Processes, and Performance

INTRODUCTION

Walmart has been successful at generating best-in-class profits over an extended period of time. Walmart's profits in 2008 were the best in the retailing sector at around $12.7 billion. eBay shook up the auction industry in the late 1990s by automating several components of the auction process. In 2008, eBay's estimated profits were about $1.8 billion. Aravind Eye Hospital, winner of the 2008 Gates Prize for Global Health, served 2,539,615 outpatients and performed 302,180 cataract surgeries between April 2009 and March 2010. Despite providing 67 percent of the outpatient visits and 75 percent of the surgeries as free service to the poor, Aravind generated healthy profits that it used to fund its growth. Netflix transformed the movie rental business from one where customers primarily visited rental stores to one where movies arrive by mail or are streamed directly to homes. In 2009, Netflix reported revenues of $1.67 billion with profits of $115 million. In contrast, Blockbuster declared bankruptcy in 2010 after many years of losing money and closed many of its movie rental stores. Each successful organization has achieved strong financial performance by providing products that meet customer expectations at a production and delivery cost that is significantly lower than the value perceived by customers. In contrast, as the Blockbuster example illustrates, inability to provide greater value to customers than the cost of production and delivery results in financial losses and the potential demise of the organization. To be successful, all organizations—software manufacturers, park districts, automakers, postal services, tax-collection agencies, and even hospitals—must provide products and services whose value to customers is much greater than the cost of production and delivery.

The financial performance of Walmart, eBay, Aravind, and Netflix illustrates how success of organizations is closely linked to their effective management of business processes that produce and deliver goods and services to their customers. Walmart's purchasing and distribution operations result in a high availability of products where and when needed, at low cost. eBay allows a seller to set up an auction in minutes (instead of weeks or months that an auction house might take) at a fraction of the cost of an auction house. Aravind has designed processes that result in high utilization of its expensive resources, allowing the company to give away free surgeries to the poor while still earning a profit. Netflix allows customers to view a wide variety of movies at a fraction of what video rental stores charged for this service. In this book, we focus on how organizations can design and manage their business processes to provide a much higher level of value to customers compared to the cost of production and delivery.

How can we represent organizations as a collection of business processes? What types of metrics do they use to monitor and manage process performance? How do organizations categorize customer expectations they seek to fulfill? How do they design their processes to deliver superior financial performance? What other enablers should organizations have in place to support the success of business processes?

This chapter provides a framework for answering these questions. In Section 1.1, we discuss how every organization can be viewed as a process. In Section 1.2, we describe performance measures that help managers to evaluate processes. In particular, we discuss financial measures, external or customer-focused measures, and internal or operational measures of performance. In Section 1.3, we look at four product attributes that determine the value that customers place on the process output: product cost, delivery-response time, variety, and quality. In Section 1.4, we study the corresponding process competencies that managers can control: processing cost, flow time, resource flexibility, and process quality. Section 1.5 defines process design, planning, control, and improvement decisions that are discussed in detail in the rest of the book. Section 1.6 discusses different process architectures that are implemented in practice.

1.1 THE PROCESS VIEW OF ORGANIZATIONS

The dictionary defines a process as "a systematic series of actions directed to some end." Building on this, we define a **process** to be *any transformation that converts inputs to outputs* (see Figure 1.1). With this definition, a single stage in an auto assembly line (say, where the seats are installed) is a process. Simultaneously, the entire assembly line is also a process that assembles components into a complete car. The process view considers any organization to be a process that consists of interconnected subprocesses. Thus, the success of any organization is determined by the performance of all its processes.

To evaluate and improve the performance of a process in terms of value created—the two key objectives of this book—we must examine the details of transformation of inputs into outputs. The following five elements of a process characterize the transformation:

1. Inputs and outputs
2. Flow units
3. Network of activities and buffers

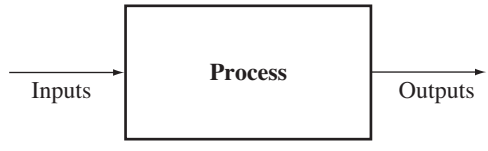


FIGURE 1.1 The Process View of an Organization (Black Box)

4. Resources
5. Information structure

Inputs and Outputs To view an organization as a process, we must first identify its inputs and outputs. **Inputs** refer to *any tangible or intangible items that "flow" into the process from the environment*; they include raw materials, component parts, energy, data, and customers in need of service. Engines, tires, and chassis are examples of inputs from the environment into an auto assembly plant. **Outputs** are *any tangible or intangible items that flow from the process back into the environment*, such as finished products, pollution, processed information, or satisfied customers. For example, cars leave as output from an assembly plant to the dealerships. So, an organization's inputs and outputs shape its interaction with its environment.

As inputs flow through the process they are transformed and exit as outputs. For example, raw materials flow through a manufacturing process and exit as finished goods. Similarly, data flows through an accounting process and exits as financial statements, and invoiced dollars (accounts receivable) flow through a billing and collection process and exit as collected dollars (cash).

Flow Units The second step in establishing a process view is obtaining a clear understanding of the **flow units**—*the item*—being analyzed. Depending on the process, the flow unit may be a unit of input, such as a customer order, or a unit of output, such as a finished product. The flow unit can also be the financial value of the input or output. For example, the flow at an Amazon warehouse can be analyzed in terms of books, customer orders, or dollars. Determining what the flow units are is important in process analysis and performance evaluation. Table 1.1 lists some generic business processes and identifies the flow units that move through the input–output transformation.

Network of Activities and Buffers The third step in adopting the process view is describing the process as a network of activities and buffers.

An **activity** is *the simplest form of transformation; it is the building block of a process*. An activity is actually a miniprocess in itself, but for our purposes of process evaluation and improvement, we are not concerned with the details of any specific activity, so a black box view of the activities will suffice. For example, when studying an interorganizational

Table 1.1 Examples of Generic Business Processes

Process	Flow Unit	Input–Output Transformation
Order fulfillment	Orders	From the receipt of an order to the delivery of the product
Production	Products	From the receipt of raw materials to the completion of the finished product
Outbound logistics	Products	From the end of production to the delivery of the product to the customer
Shipping	Products/ Orders	From the shipment of the product/order to the delivery to the customer
Supply cycle	Supplies	From issuing of a purchase order to the receipt of the supplies
Customer service	Customers	From the arrival of a customer to their departure
New product development	Projects	From the recognition of a need to the launching of a product
Cash cycle	Cash	From the expenditure of funds (costs) to the collection of revenues

process such as a supply chain that includes suppliers, manufacturers, distributors, and retailers, it is enough to view each organization as one activity or black box rather than looking at all the activities that take place within each organization. When studying each organization more fully, however, we must study its particular transformation process by looking closely at its specific activities. At this level, we would look at activities such as spot welding sheet metal to an auto chassis, checking in passengers at airport terminals, entering cost data in accounting information systems, and receiving electronic funds transfers at collection agencies.

A **buffer** *stores flow units that have finished with one activity but are waiting for the next activity to start*. For example, a patient who has registered at an emergency room waits in a waiting room to see the doctor. Similarly, cars that have been painted wait in a storage area before entering the assembly line. For physical goods, a buffer often corresponds to a physical location where the goods are stored. A process, however, may also have a buffer that does not correspond to any physical location, as in case of customer orders waiting to be processed. You can think of storage in a buffer as a special activity that transforms the time dimension of a flow unit by delaying it. In business processes, storage is called **inventory** which is *the total number of flow units present within process boundaries*. The amount of inventory in the system is an important performance measure that we discuss in detail in Chapter 3.

Process activities are linked so that the output of one becomes an input into another, often through an intermediate buffer—hence the term a **network of activities and buffers**. This network describes the specific **precedence relationships** among activities—*the sequential relationships that determine which activity must be finished before another can begin*. As we will see later, the precedence relationships in a network structure strongly influence the time performance of the process. In multiproduct organizations, producing each product requires activities with a specific set of precedence relationships. Each network of activities, then, can have multiple "routes," each of which indicates precedence relationships for a specific product.

Figure 1.2 shows a process as a network of activities and buffers and the routes or precedence relationships among them.

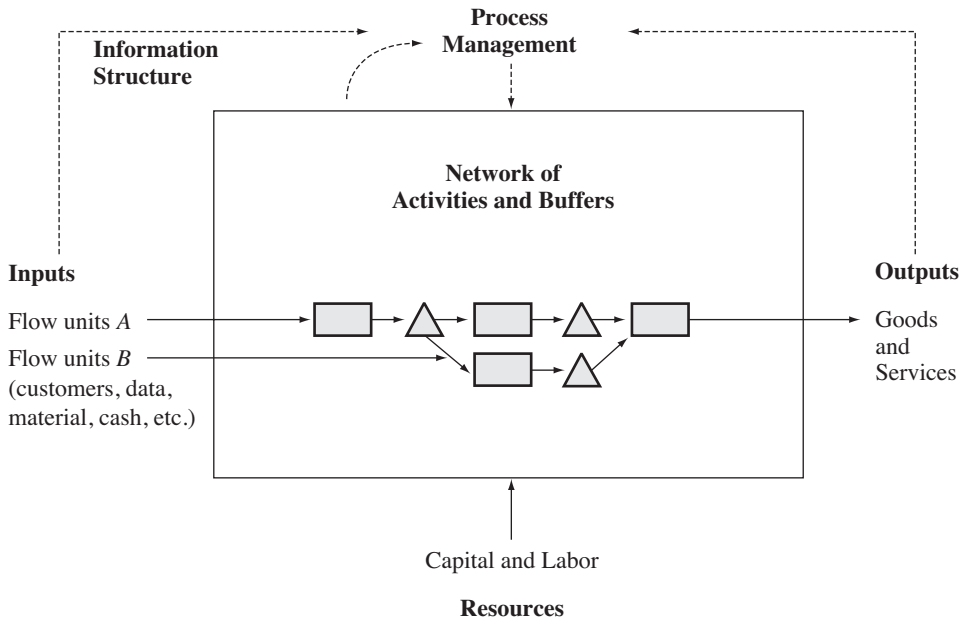


FIGURE 1.2 A Process as a Network of Activities and Buffers

Resources The fourth element of the process view consists of organizational resources. From an operations perspective, **resources** are *tangible assets that are usually divided into two categories*:

- **Capital**—*fixed assets such as land, buildings, facilities, equipment, machines, and information systems*
- **Labor**—*people such as engineers, operators, customer-service representatives, and sales staff*

Resources facilitate the transformation of inputs into outputs during the process. Some activities require multiple resources (a welding activity, for instance, requires a worker and a welding gun), and some resources can perform multiple activities (some workers can not only weld but also drill). The allocation of resources to activities is an important decision in managing any process.

Information Structure The fifth and final element of the process view of an organization is its **information structure**, which *shows what information is needed and is available to whom in order to perform activities or make managerial decisions*.

Thus, we can now define a **business process** as *a network of activities separated by buffers and performed by resources that transform inputs into outputs*. **Process design** *specifies the structure of a business process in terms of inputs, outputs, the network of activities and buffers, and the resources used*. **Process flow management**, therefore, is *a set of managerial policies that specify how a process should be operated over time and which resources should be allocated to which activities*. This process view of organizations will be our basis for evaluating and improving organizational performance. We will see how process design and process flow management significantly affect the performance of every organization.

The process view is applicable to a variety of organizations. It can represent a manufacturing process that transforms raw materials into finished products as well as services performed by accounts receivable departments, product design teams, computer rental companies, and hospitals. The process view is also a convenient tool for representing cross-functional processes within an organization, including production, finance, and marketing-related functions and supplier relationships. By incorporating buffers, we also account for handoffs or interfaces between different people or activities—typically the areas where most improvements can be made. In addition, the process view can be adopted at a very broad level, such as the supply chain, or at a very micro level, such as a workstation in a plant. The process view is "customer-aware"—it always includes the customer—the person who receives the outputs.

Value stream mapping or **value chain mapping** is *a tool used to map the network of activities and buffers in a process identifying the activities that add value and those like waiting that are wasteful*. The goal of value stream mapping is to enable process designers and managers to focus on process improvement by adding value to the final product. Processes should ideally be designed and managed to add value at every step of the process. This book is aimed at identifying design and management strategies that improve the value adding potential of a process.

The process view highlights the fact that every organization is a collection of interconnected processes. For example, filling a customer order at Amazon involves the order-taking process, the picking and packing process at the warehouse, and the shipping process. The performance of the picking and packing process relies on accurate receiving and replenishment of material at the warehouse. If an item has been misplaced during receiving, the picker may not be able to find it when looking to fill an order. The basic point to understand is that the success of any organization requires alignment of effort across all its processes.

Aravind Eye Hospital provides an excellent example of interconnected processes designed to provide high quality eye care at affordable prices. Bright sunlight and a

genetic predisposition make people in India particularly vulnerable to cataracts. Millions of Indians lose their sight by their fifth decade. The poor who are unable to access or afford treatment needlessly go blind. Aravind was founded by Dr. Venkataswamy, an ophthalmologist, with a mission of eradicating needless blindness in India. Success required both affordability and improved access to eye care.

Aravind started by treating paying patients and using the profits to offer free care to those who could not afford it. While this made it affordable, the service was often not accessible to patients who could not afford transportation and required a relative to accompany them, often at the expense of several days' lost income. So, Aravind doctors visited villages offering eye camps and the organization added its own buses and a group of ophthalmic assistants who accompanied patients every step of the way from first examination, through surgery to the bus ride back home. To further improve access at low cost, Aravind opened vision care centers in rural areas where trained paramedical staff worked remotely with doctors at the Madurai hospital to offer diagnosis and treatment. To keep costs low, Aravind ensures high utilization of its expensive resources—doctors and surgical equipment. Surgical equipment is used all day and doctors only focus on performing surgery with preoperative and postoperative care largely handled by nurses. The level of quality has consistently been high enough to attract a significant number of paying patients. The process view of an organization allowed Aravind to constantly identify changes that created greater value for the patient while keeping costs under control. As a result, Aravind today is the world's largest provider of eye care services treating a large fraction of its patients for free.

The process view of organizations is our main tool for the following:

1. Evaluating processes
2. Studying the ways in which processes can be designed, restructured, and managed to improve performance

1.2 PERFORMANCE MEASURES

What determines the effectiveness of a process? Any reasonable answer to this question must be based on two factors:

1. Evaluation and measurement of the firm's current and past performance
2. Future goals as expressed by the firm's strategy

In order to assess and improve the performance of a business process, we must measure it in quantifiable terms. In this section, we identify several quantifiable measures of process performance—financial, external, and internal.

1.2.1 The Importance of Measurement: Management by Fact

Longtime General Motors chairman Alfred Sloan defined a "professional manager" as someone who manages by fact rather than by intuition or emotion. By capturing facts in an objective, concrete, and quantifiable way, we get a clear picture of the relationship between controllable process competencies and desired product attributes and thus are able to set appropriate performance standards (see Table 1.2). Performance measurement is essential in designing and implementing incentives for improving products and processes and for assessing the result of our improvements.

1.2.2 Types of Measures: Financial, External, and Internal

The financial performance of a process is based on the difference between the value that outputs of the process (products and services) provide to customers and their

Table 1.2 The Importance of Measurement

"When you can measure what you are speaking about, and express it in numbers, you know something about it."

—Lord Kelvin (1824–1907)

"Count what is countable, measure what is measurable, and what is not measurable, make measurable."

—Galileo Galilei (1564–1642)

"Data! Data! Data! I can't make bricks without clay."

—Sherlock Holmes in *The Adventure of Copper Beeches* by Sir Arthur Conan Doyle (1859–1930)

"In God we trust, everyone else must bring data."

—W. Edwards Deming (1900–1993)

cost of production and delivery. Solid financial performance depends on the ability of a process to effectively meet customer expectations. Thus, process management requires external measures that track customer expectations and internal measures that gauge the effectiveness of the process in meeting them. External measures indicate how customers view the organization's products and services, whereas internal measures identify areas where the process is performing well and areas where improvement is necessary.

FINANCIAL MEASURES

Financial measures track the difference between the value provided to customers and the cost of producing and delivering the product or service. The goal of every organization is to maximize this difference. Profit-making enterprises aim to keep part of this difference to themselves as profit by charging appropriate prices for their goods and services. Not-for-profit organizations generally leave a large part of the difference with their clients using the rest to maintain viability and grow. In either case, increasing the difference between the value provided to customers and the cost of production and delivery is a key goal.

Each quarter, most organizations report three types of financial measures to shareholders and other stakeholders:

1. Absolute performance (revenues, costs, net income, profit)
2. Performance relative to asset utilization (accounting ratios such as return on assets, return on investment, and inventory turns)
3. "Survival" strength (cash flow)

Although the ultimate judge of process performance, financial measures are inherently lagging, aggregate, and more results than action oriented. They represent the goal of the organization but cannot be used as the sole measures to manage and control processes. Managing and controlling a process based only on financial measures would be like driving a car while looking in the rear-view mirror. Thus, it is important to link financial measures to external measures that track customer satisfaction with the process output and internal measures that track operational effectiveness.

EXTERNAL MEASURES

To improve its financial performance, a firm must attract and retain customers by providing goods and services that meet or exceed their expectations. Customer expectations can be defined in terms of four critical attributes of the process output (products and services) cost, response time, variety, and quality. For example, FedEx customers

expect speed and reliability and are willing to pay more for it than, say, customers of the U.S. Postal Service, who expect a lower cost and are willing to tolerate a longer delivery time. A person buying a Lexus expects a high quality ride, a range of options, and very responsive service. A person buying a Toyota Corolla, in contrast, expects to pay a much lower price and is willing to compromise on features, options, and responsiveness of service. Customer satisfaction is then linked to whether the performance of the product along the four attributes meets or exceeds customer expectations.

External measures track customer expectations in terms of product or service cost, response time, variety, and quality as well as customer satisfaction with performance along these dimensions. External measures can be used to estimate the value of goods or services to customers. For example, the American Society for Quality and the University of Michigan has developed the American Customer Satisfaction Index, which tracks overall customer satisfaction in several manufacturing and service industries and public sectors. Each score is a weighted average of customer responses to questions relating to perceptions of service, quality, value, and the extent to which products meet expectations. J. D. Power and Associates and Consumer Reports also provide surveys of customer satisfaction and rankings of products and services including automobiles, appliances, and hotels and restaurants.

Measures that track customer dissatisfaction with a product are also good external measures that can help guide future improvement. Number of warranty repairs, product recalls, and field failures are some measurable signs of potential customer dissatisfaction. Although number of customer complaints received is a direct measure of customer dissatisfaction, research shows that only about 4 percent of dissatisfied customers bother to complain. Those who do complain, therefore, represent just the tip of the iceberg. Customer dissatisfaction decreases customer retention, leading to lower revenues and increased costs in the long run. It is estimated that organizations typically lose 20 percent of their unsatisfied customers forever and that the cost of attracting a new customer is about five times that of serving a current customer.

Customer-satisfaction measures represent an external market perspective that is objective and bottom-line oriented because it identifies competitive benchmarks at which the **process manager**—*the person who plans and controls the process*—can aim. However, they measure customer satisfaction at an aggregate, not at an individual customer, level. They are also more results oriented than action oriented: In other words, they cannot indicate *how* the manager might improve processes. Finally, they are lagging rather than leading indicators of success, as they are "after-the-fact" assessments of performance. To be operationally useful, they must be linked to internal measures that the process manager can control.

INTERNAL MEASURES

A process manager does not directly control either customer satisfaction or financial performance. In order to meet customer expectations and improve financial performance, a manager requires internal operational measures that are detailed, that can be controlled, and that ultimately correlate with product and financial performance. Customer expectations in terms of product or service cost, response time, variety, and quality can be translated into internal measures that track the performance of the process in terms of processing cost, flow time, process flexibility, and output quality. Internal performance measures can thus be a predictor of external measures of customer (dis)satisfaction (and thus financial performance), if customer expectations have been identified accurately.

For example, an airline's on-time performance may be translated into the following internal goal: "Average arrival and departure delays should not exceed 15 minutes." The responsiveness of its reservation system could be measured by "the time

taken to answer the telephone," with a specified goal of "30 seconds or less 95 percent of the time." Similarly, waiting time for service at a bank teller's window or for registration and admission at a hospital can all be measured, monitored, and standardized. Likewise, product availability at a retailer can be measured by such standards as "fraction of demand filled from the stock on hand" or "average number of stockouts per year." At a call center, service (un)availability can be measured by the proportion of customer calls that must wait because "all operators are busy." The goal might be to reduce that proportion to "no more than 20 percent." At an electric utility company, service availability might be measured by "frequency or duration of power outages per year" with a target of "no more than 30 minutes per year." In each case, the internal measure is an indicator of how satisfied the customer is likely to be with process performance.

Process flexibility can be measured either by the time or cost needed to switch production from one type of product or service to another or by the number of different products and services that can be produced and delivered.

When measuring product quality, managers must be specific as to which of the many quality dimensions they are concerned with: product features, performance, reliability, serviceability, aesthetics, and conformance to customer expectations. Reliability, for example, is measured in terms of durability and frequency of repair and can be assessed by technical measures like the following:

- Failure rate, which measures the probability of product failure
- Mean time between failures (MTBF), which indicates how long a product is likely to perform satisfactorily before needing repair

Serviceability can be measured using mean time to repair (MTTR), which indicates how long a product is likely to be out of service while under repair.

Although customers can readily identify product features, performance can be assessed only through actual experience relative to expectations. For example, the primary features offered at McDonald's are reliability and speed from a low-priced, limited menu. In contrast, dining at a high-end restaurant may involve highly customized dishes that change with the availability of fresh produce and the dining experience may be expensive and may take hours. In each case, performance is judged based on customer expectations. Whereas a McDonald's customer may be unhappy if it takes a long time, a customer at a high end restaurant typically prefers a more relaxed experience that is not hurried.

Knowing the external product measures expected by customers, the process manager must translate them into appropriate internal process measures that affect external measures. In order to be effective, internal measures must meet two conditions:

1. They must be linked to external measures that customers deem important.
2. They must be directly controllable by the process manager.

Obviously, measuring and improving a feature that the customer does not value is a waste of time and resources. Moreover, if we do not know how an internal process variable affects a product measure, we cannot control it. Ignoring one or both of these conditions has sabotaged many process improvement programs (see Kordupleski et al., 1993).

1.3 PRODUCTS AND PRODUCT ATTRIBUTES

Products are *the desired set of process outputs*. (The process may also produce by-products, such as heat, pollution, or scrap, which are not desired by or delivered to customers.) Given that products may be physical goods, services performed, or a combination of both, there are some differences (and many similarities) between goods and

services that managers must consider when designing or managing the processes that deliver them. One difference is that unlike tangible goods, services include tangible and intangible aspects experienced by the customer, such as being treated by a doctor or receiving investment advice. Another difference is that some services, such as a haircut, are often produced and consumed simultaneously and cannot be produced in advance, whereas physical goods can be produced and stored for later consumption.

Different customers may have different expectations of a specific product. If the process is to produce and deliver products that satisfy all customers, the process manager must know what these expectations are. External measures help a process manager identify key **product attributes**—*those properties that customers consider important*—that define customer expectations. For example, external measures help FedEx define customer expectations in terms of delivery time, reliability, and price for its next-day-delivery product. We categorize product attributes along the following four dimensions:

1. **Product cost** is *the total cost that a customer incurs in order to own and experience the product*. It includes the purchase price plus any costs incurred during the lifetime of the product, such as costs for service, maintenance, insurance, and even final disposal. Cost is important because customers usually make purchase decisions within budget constraints.
2. **Product delivery-response time** is *the total time that a customer must wait for, before receiving a product for which he or she has expressed a need to the provider*. Response time is closely related to product availability and accessibility. If a manufactured good is on store shelves, the response time is effectively zero. If it is stocked in a warehouse or a distribution center, response time consists of the transportation time needed to get it to the customer. If a firm does not stock the product and produces only to order, response time will also include the time required to produce the product.

   With services, response time is determined by the availability of resources required to serve the customer. If resources are not immediately available, the customer must wait. For example, in the absence of an idle checkout clerk, a customer has to wait in a checkout line for service. Generally, customers prefer short response times, as immediate gratification of needs is typically preferred over delayed gratification. In many instances the reliability of the response time is at least as important as its duration.
3. **Product variety** is *the range of choices offered to the customer to meet his or her needs*. Variety can be interpreted and measured at different levels. At the lowest level, we can measure variety in terms of the level of customization offered for a product. This includes options offered for a particular car model or the number of colors and sizes for a style of jeans. At a higher level, variety can be measured in terms of the number of product lines or families offered by a firm. For example, a car manufacturer like General Motors offering a full range of automobiles like compacts, sports cars, luxury sedans, and sport-utility vehicles (SUVs) provides a greater variety than a manufacturer like Ferrari that offers only sports cars. Similarly, a retail store offering the full range of apparel from casual to business to formal wear offers more variety than a store focused on providing only tuxedoes. Whereas standard, commodity products have little variety, custom products may be one-of-a-kind items tailored specifically to customers' unique needs or wishes. For example, when purchasing apparel in a department store, customers must choose from a limited selection. In contrast, when ordering a suit at a custom tailor, each customer can provide different specifications that meet personal needs and desires that constitute, in effect, an almost endless range of product variety.
4. **Product quality** is *the degree of excellence that determines how well the product performs*. Product quality is a function of effective design as well as production

Table 1.3 What Is Quality?

"Quality is recognized by a non-thinking process, and therefore cannot be defined!" —R. M. Pirsig in *Zen and the Art of Motorcycle Maintenance*
"That which makes anything such as it is." —*Funk and Wagnall's Dictionary*
"Fitness for use." —J. Juran and American Society of Quality Control
"Conformance to requirements." —P. Crosby
"Closeness to target—deviations mean loss to the society." —G. Taguchi
"Total Quality Control provides full customer satisfaction at the most economical levels." —A. Feigenbaum
"Eight dimensions of quality are: Performance, Features, Conformance, Reliability, Serviceability, Durability, Aesthetics, and Perception." —D. Garvin

that conforms to the design. It may refer to tangible, intangible, and even transcendental characteristics of product experience. Product quality is often the most difficult product attribute to define and measure because subjective judgment and perception play important roles in a customer's assessment of quality. Table 1.3 lists some definitions of quality. Each definition tries to capture something from elusive to all-inclusive, largely because quality must be seen from both the customer's and the producer's perspectives.

From the customer's perspective, quality depends on a product's features (what it can do), performance (how well it functions), reliability (how consistently it functions over time), serviceability (how quickly it can be restored), aesthetics, and conformance to expectations. Whereas product features and performance are influenced by quality of design, reliability is more heavily influenced by how well the production process conforms to the design. The styling, size, options, and engine rating of an automobile are its features. Acceleration, emergency handling, ride comfort, safety, and fuel efficiency are aspects of performance, while durability and failure-free performance over time represent its reliability.

A product may be defined as a bundle of the four attributes—cost, time, variety, and quality. When these four attributes are measured and quantified, we can represent a product by a point in the associated four-dimensional matrix, or *product space*, of cost, time, variety, and quality. (The "product space" image will be a useful metaphor to define strategy in Chapter 2.) Well-defined external measures track product performance along these four dimensions, relative to the competition and relative to customer expectations.

The value of a product to the customer is measured by the utility (in economic terms) that he or she derives from buying the combination of these attributes. In general, high-quality products, available in a wide variety, delivered quickly and at a low cost provide high value to the customers. Product value or utility is a complex function of the four product attributes. It may be easy to define qualitatively, but it is difficult to measure in practice. A reasonable estimate of **product value** *is the maximum price that a specific customer is willing to pay for a product*. Of course, this willingness to pay varies from customer to customer, giving rise to the familiar relationship between price and demand described by economists.

Customers prefer products with all the attributes—they want products and services that are good, fast, and inexpensive. However, producing and delivering all products involves trade-offs: Some products are good quality but not delivered as fast; some are inexpensive but not as good quality. When customer expectations depend on the availability of competing products, an important strategic business decision involves selecting the right combination of product attributes that will result in a product that appeals to a particular segment of the market, as we will see in Chapter 2. Moreover, to keep abreast of the competition, there must be continuous improvement in product variety and quality and a decrease in cost and delivery-response time.

1.4 PROCESSES AND PROCESS COMPETENCIES

Processes produce and deliver products by transforming inputs into outputs by means of capital and labor resources. *The process of producing physical goods is typically called* **manufacturing**. *Processes that perform services are called* **service operations**. We refer to *business processes that design, produce, and deliver goods and services* simply as **operations**. Given the many similarities, we highlight some of the unique aspects of service operations. Many service operations, such as a hospital, require the physical presence of the customer who undergoes or participates in at least part of the process. This introduces variability from one customer to the next and increases the importance of factors such as the attractiveness of the process environment and friendliness of the labor resources. (The term "backroom operations" refers to those aspects of service operations that are hidden from customers.) Services involve significant interaction with the customer and are often produced and consumed simultaneously, which makes it harder to identify internal measures of performance that could be leading indicators of external customer satisfaction.

A process manager aims to improve financial performance by effectively producing products that satisfy customer expectations in terms of the four product attributes—cost, response time, variety, and quality. In this section, we use four dimensions for measuring the competence of processes to produce and deliver the corresponding four product attributes:

1. **Process cost** is *the total cost incurred in producing and delivering outputs*. It includes the cost of raw materials and both the fixed and the variable costs of operating the process. (For our purposes, this is as specific as we need to be about the ways accounting practices allocate costs to time periods and products.)
2. **Process flow time** is *the total time needed to transform a flow unit from input into output*. It includes the actual processing time as well as any waiting time a flow unit spends in buffers. Process flow time depends on several factors including the number of resource units as well as the speed of processing by each resource unit.
3. **Process flexibility** measures *the ability of the process to produce and deliver the desired product variety*. Process flexibility depends on the flexibility of its resources: Flexible resources (such as flexible technology and cross-trained workers or "generalists") can perform multiple different activities and produce a variety of products. Dedicated or specialized resources, in contrast, can perform only a restricted set of activities, typically those designed for one product. Another dimension of process flexibility is its ability to deal with fluctuating demand. A steel mill cannot readily alter the amount of steel it produces at a time. An auto repair shop, in contrast, finds it easier to change the number of cars repaired each day. Information technology today has made it very easy for people to customize their online experience whether viewing a newspaper or a shopping site. Online retail sites are able to customize the products shown (and in some instances the price charged) based on individual preferences (and willingness to pay).

4. **Process quality** refers to *the ability of the process to produce and deliver quality products*. It includes process accuracy (precision) in producing products that conform to design specifications, as well as reliability, and maintainability of the process.

Process competencies determine the product attributes that the process is particularly good at supplying. For example, McMaster-Carr, a distributor of materials, repair, and operations (MRO) products, has a process that has high flexibility, short flow time, and high quality. It does not, however, have a low-cost operation. Therefore, process competencies at McMaster-Carr allow it to supply a large variety of MRO products quickly and reliably, while charging its customers a premium price. Customers looking for a large quantity of a single item at low price, therefore, do not go to McMaster-Carr.

Shouldice Hospital in Canada focuses exclusively on hernia operations for otherwise healthy patients. The founder developed and standardized a repeatable surgical procedure that requires only local anesthesia and encourages patient movement, participation, and socialization through excellent ambulatory care provided in a non-hospital-like environment. The Shouldice process provides very high quality service at relatively low cost. It is, however, very inflexible and will not accept patients who have any risk factor and certainly does not treat patients for anything other than hernia. The competencies required in an emergency room process are very different from those of Shouldice. Given the wide variety of patients that an emergency room has to treat, its process competencies must include flexibility, quick response, and high quality.

1.5 ENABLING PROCESS SUCCESS

The success of a process is ultimately dependent on its ability to meet or exceed customer expectations in a cost-effective manner. To enable successful performance, companies must address the following five questions effectively:

1. What should the process design or architecture be?
2. What metrics should be used to track performance of a process?
3. What policies should govern process operations?
4. How should process performance be controlled over time?
5. How should process performance be improved?

During process design, managers select the process architecture that best develops the competencies that will meet customer expectations of the product. Process design decisions include plant location and capacity, product and process design, resource choice and investment (capital/technology and labor), and scale of operation. For example, Toyota has plants in most major markets that produce large volumes of the Corolla using an assembly line where workers stay in their position repeating a task while cars move from one station to the next. This allows Toyota to provide a large volume of Corollas quickly at a low cost but with a limited variety. In contrast, Ferrari manufactures all its cars in Italy using a much more flexible process with highly skilled and flexible workers. Even though it takes longer to produce a car (and especially one that is very expensive!), Ferrari is able to accommodate a high degree of customization. Given the small volume of sales, a single plant allows Ferrari to achieve some economies of scale relative to having plants in every major market.

During **metric identification**, *managers identify measurable dimensions along which the performance of the process will be tracked.* Ideally, **process metrics** are derived from customer expectations and a company's strategic goals, which should relate to desired process competencies. For example, a company focused on being responsive must carefully track its order fulfillment time to ensure on time delivery. A company focused on product innovation should track the percentage of revenue derived from products

introduced over the past 12 months. The goal is to provide managers with information about performance that allows them to plan, control, and improve the process to better meet customer expectations about the product.

Process planning is *identifying targets for the various metrics and specifying managerial policies that support the achievement of these targets*. Managerial policies specify the operation of the process and use of resources over time to best meet customer demand. At a retail store, managerial policies specify how many units of each product should be carried in stock and when replenishment orders should be placed to ensure a desired level of product availability. At a call center, managerial policies specify the number of customer service representatives that should be available by day of week and time of day to keep response time under a desired level. At an apparel retailer like Zara, managerial policies specify when supply should come from a responsive but high-cost source in Europe and when it should come from a low-cost but longer lead-time source in Asia. In each instance, managerial policies aim to provide a targeted level of performance in terms of cost, time, flexibility, and quality.

Process control is *the tactical aspect of process management that is focused on continually ensuring that in the short run, the actual process performance conforms to the planned performance*. Every process is subject to variation, partly because of design limitations and partly because external factors may intrude into the process environment. A machine filling cereal boxes will have some natural variation in the quantity of cereal filled in each box. A malfunction of the filling system (an external factor intruding onto the process), however, may result in consistent over or under filling. The objective of process control is to continuously monitor process performance to identify instances where external factors may have intruded into the process environment, limiting its ability to conform to the planned performance. In such instances corrective action can be taken to bring process performance back to the planned level. Control decisions include monitoring and correcting product cost, delivery time, inventory levels, and quality defects.

For **process improvement**, *managers identify metrics that need to be improved in the long run and work on changes in process design or planning that are required to achieve this improvement*. For example, Toyota has identified that it is important for its suppliers to be more flexible, operate with lower inventories, and be able to respond to a Toyota order quickly. Toyota, therefore, works with its suppliers to reduce changeover times in all production lines and detect defects as soon as they are introduced. These process changes improve its performance in terms of flexibility, inventories, and time.

Given that in an organization no process exists in isolation (and no organization exists in isolation), for optimal performance it is important that each of the five questions identified earlier be addressed at both an interprocess and interorganization level. For example, it is not enough for Netflix to simply track when a movie leaves its warehouse because a customer only cares about when he or she receives the movie, not when Netflix shipped it. For optimal performance, Netflix should answer all five questions to align well with the U.S. Postal Service and improve the overall process of getting a movie to a customer on time. For movies that are delivered online, it is not enough for Netflix to ensure that it has enough server capacity of its own. To ensure optimal viewing experience, Netflix has to account for constraints along the way from their servers to the viewer's home.

1.6 SOME BASIC PROCESS ARCHITECTURES

Process architecture is defined by *the types of resources used to perform the activities and their physical layout in the processing network*. Automobile assembly plants have process architecture with specialized resources laid out in a rigid sequence that is common for

all cars produced. A tool and die shop, in contrast, has process architecture that uses flexible resources to work on a variety of products, each with a different sequence of activities. An ambulatory cancer care center has a process architecture that is standardized, with every patient going through a similar set of steps (perhaps with different medication) to complete treatment. At an emergency room, in contrast, the process architecture needs to be flexible enough to deal with a wide variety of patient needs. Process competencies are strongly influenced by the process architecture. Our ultimate goal is to design process architecture with competencies that align well with customer expectations, as will be discussed in Chapter 2.

Most process architectures fall somewhere on the spectrum between two extremes: a flexible job shop process and a specialized flow shop process.

JOB SHOPS

At one extreme, a **job shop** uses *flexible resources to produce low volumes of highly customized, variety products*. Examples of job shops include artisan bakeries, tool and die shops, management consulting firms, law firms, and architectural and design companies. Job shops use general-purpose resources that can perform many different activities and locate similar resources together. This design is called a **functional layout** or **process layout** because it *groups organizational resources by processing activities or "functions" in "departments."* For example, a job shop manufacturing process (such as a tool and die or machine shop) groups all its presses together in a stamping department and all its mills in a milling department.

A job shop usually has many products simultaneously flowing through the process, each with its own resource needs and route. Therefore, it is often more practical to represent a job shop as a network of resources instead of a network of activities. In a network of resources, rectangular boxes represent resources that are grouped into departments (such as an X-ray, accounts payable, or stamping department) instead of activities. The flowchart on the left in Figure 1.3 shows an example of the functional layout of a process with four resource groups (labeled A, B, C, and D) that produces two products. Resources A, D, and B perform activities on product 1, while product 2 calls for resources C, B, and A. The set of activities for each product is now assigned to the resources with the routes representing the precedence relationships, as before.

Because the sequence of activities required to process each product (i.e., the routes) varies from one job to the next, job shops typically display jumbled work flows with large amounts of storage buffers and substantial waiting between activities. To direct the work flow, job shops typically use highly structured information systems.

Because of the high variety of products flowing through the job shop, resources often need setups before they can be changed over from the activities required for one product to those required for another. This changeover results in delays, loss of production

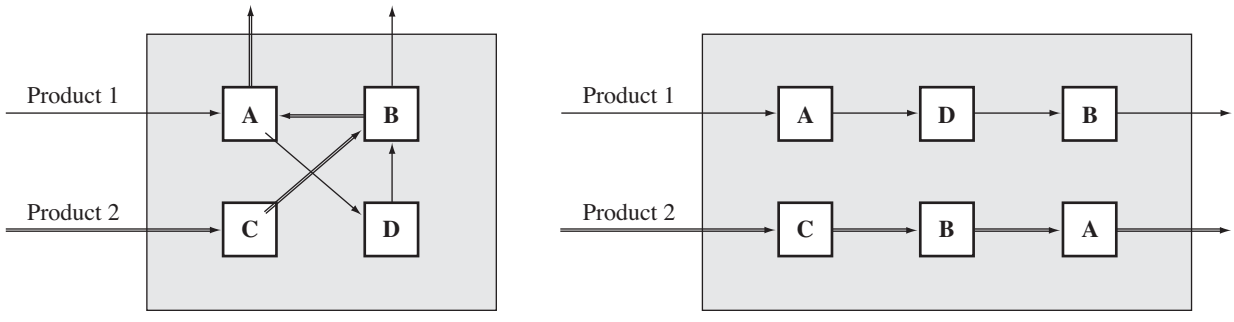


FIGURE 1.3 Functional Layout (left) versus Product Layout (right)

and a fluctuating workload. In terms of process competencies, a job shop typically has high process flexibility that permits product customization but has high processing costs and long flow times.

FLOW SHOPS

At the other extreme, a **flow shop** uses *specialized resources that perform limited tasks but do so with high precision and speed*. The result is a standardized product produced quickly in large volumes. Because of the specialized resources and expertise developed by workers through repetition, product quality tends to be more consistent. Although the high-processing capacity needed to produce large volumes entails high fixed costs for plant and equipment, these costs are spread over larger volumes, often resulting in the low variable processing cost that characterizes economies of scale. Resources are arranged according to the sequence of activities needed to produce a particular product, and limited storage space is used between activities. *Because the location of resources is dictated by the processing requirements of the product, the resulting network layout is called a* **product layout**. The flowchart on the right in Figure 1.3 shows the two-product process in product layout. Each product is now produced on its own "production line" with product-dedicated resources. Notice that dedicating resources to a product may necessitate duplication (and investment) of a resource pool, such as for resources A and B in Figure 1.3. On the positive side, limiting the product variety allows specialization of dedicated resources. Therefore, flow shops typically have shorter process flow time than job shops.

The most famous example of a flow shop is the automobile assembly line pioneered by Henry Ford in 1913. An assembly line is actually an example of a discrete flow shop: Products are produced part by part. In contrast, beverage companies, steel plants, oil refineries, and chemical plants are continuous flow shops: Sometimes called "processing plants," they produce outputs, such as beer, steel, and oil, in a continuous fashion. Although the rigid layout of resources and their highly specialized nature prevent the process from providing significant product variety, the flow shop remains the crown jewel of the industrial revolution—its hallmark is low unit-processing cost, short flow time, and consistent quality at high volumes.

All real-world processes fall somewhere along the spectrum between these two extremes. In the early stages of a product's life cycle, for example, because volumes are low and uncertain, high capital investment in specialized resources cannot be justified. Consequently, a flexible process like a job shop is appropriate. As the product matures, however, volume increases and product consistency, cost, and response time become critical. The flow shop then becomes the more appropriate process architecture. At the end of the product life cycle, volumes have declined, and perhaps only replacements are needed. Again, the job shop becomes more attractive.

Designing the appropriate process architecture is perhaps the most important decision an organization can make because the architecture dictates what a process will be good at and what it should focus on. Success ultimately results from designing a process which is good along precisely the dimensions that are valued most by the customers.

1.7 THE PLAN OF THE BOOK

The remainder of this book focuses on the management of operations in general and process flows in particular. Chapter 2 examines strategic decisions about product attributes and matching process competencies to these product decisions. In Part II, we analyze key process performance measures in detail. Part III is devoted to process planning and control and stresses the means by which managers achieve desired performance in the presence of uncertainty. Part IV concludes with the principles of process synchronization, integration, and improvement.

Summary

A process is a network of activities and buffers that uses resources to transform inputs into outputs. Any organization or any part of an organization can be viewed as a process. The effectiveness of a process is ultimately determined by its financial performance—the difference between the value provided to customers and the cost of producing and delivering the product. Any financial measure, however, is a lagging measure of performance and thus cannot be used to manage and control the process.

To improve financial performance, a firm must attract and retain customers by providing goods and services that meet or exceed their expectations. Customer expectations are defined in terms of four key product attributes—cost, delivery-response time, variety, and quality. From a customer's perspective, a product is thus a bundle of these four attributes. The value of a product is measured by the utility that the customer derives from buying the combination of these attributes. To improve financial performance, a firm must identify and deliver attributes that are valued by customers at a lower cost than the value delivered.

Product attributes are the output of a process and can be measured only after the processing is complete. As leading indicators of performance, the manager must internally manage the process competency in terms of cost, flow time, flexibility, and quality. The competencies of a process determine the products that the process will be particularly good at supplying. Different process architectures result in different process competencies. At one extreme, a job shop has high process flexibility that permits product customization but at high processing costs and long flow times. At the other extreme, a flow shop provides low cost, short flow times, and consistent quality but cannot produce a wide variety of products.

Key Terms

- Activity
- Buffer
- Business process
- Capital
- Flow shop
- Flow units
- Functional layout
- Information structure
- Inputs
- Inventory
- Job shop
- Labor
- Manufacturing
- Metric identification
- Network of activities and buffers
- Operations
- Outputs
- Precedence relationships
- Process
- Process architecture
- Process control
- Process cost
- Process design
- Process flexibility
- Process flow management
- Process flow time
- Process improvement
- Process layout
- Process manager
- Process metrics
- Process planning
- Process quality
- Product attributes
- Product cost
- Product delivery-response time
- Product layout
- Product quality
- Product value
- Product variety
- Products
- Resources
- Service operations
- Value chain mapping
- Value stream mapping

Discussion Questions

1.1 Several examples of organizations are listed here. For each, identify underlying business processes in terms of inputs, outputs, and resources employed. What financial, external, and internal performance measures should each organization use? Who are their customers, and what product attributes do they consider important? What process competencies should each organization aim for?

- Personal computer manufacturer
- Telephone company
- Major business school
- Hospital

- Federal penitentiary
- Red Cross
- Law firm
- Fast-food restaurant
- Inner-city school
- Local bank
- Art museum
- Public park
- Toothpaste manufacturer

1.2 Compare Walmart and a convenience store like 7-Eleven in terms of the product attributes that customers expect. What process competencies does each organization aim to develop?

1.3 Compare McDonald's and a fine dining restaurant in terms of the product attributes that customers expect. What process competencies should each organization aim to develop? Would you expect each process to be more like a job shop or a flow shop? Why?

1.4 As a product moves through its life cycle from introduction to maturity to decline, how do the attributes that customers consider important change? What are the implications in terms of the process competencies that need to be developed? What process type is appropriate in the introductory phase? In the maturity phase? Why?

Selected Bibliography

Anupindi, R. M. Aundhe and M. Sarkar. 2009. "Healthcare Delivery Models and the Role of Telemedicine." In *Indian Economic Superpower: Fiction or Future.* Singapore: World Scientific.

Hammer, M. "Deep Change: How Operational Innovation Can Transform Your Company." *Harvard Business Review* 82, no. 4 (April 2004): 84–93.

Hammer, M. "The Process Audit." *Harvard Business Review* 85, no. 4 (April 2007): 111–123.

Kordupleski, R. E., R. T. Rust, and A. J. Zahorik. "Why Improving Quality Doesn't Improve Quality (or Whatever Happened to Marketing?)." *California Management Review* 35, no. 3 (Spring 1993): 82–95.

Leavitt, T. "The Industrialization of Service." *Harvard Business Review* 54, no. 5 (September–October 1976): 63–74.

Sam Lazaro, F. "Two Decades On, India Eye Clinic Maintains Innovative Mission." *PBS Newshour* (September 2, 2009). Accessed on April 18, 2011 at www.pbs.org/newshour/updates/health/july-dec09/eye_09-02.html. Also see video available at www.aravind.org/PBS_newshour.html.

Schmenner, R. W. *Plant and Service Tours in Operations Management*. Upper Saddle River, N.J.: Prentice Hall, 1997.

Schmenner, R. W. *Service Operations Management*. Upper Saddle River, N.J.: Prentice Hall, 1995.

Wheelwright, S. C. "Reflecting Corporate Strategy in Manufacturing Decisions." *Business Horizons* 21 (February 1978): 57–66.

Chapter

2

Operations Strategy and Management

INTRODUCTION

With regards to the airline industry, Warren Buffett has famously been quoted as saying that "if a farsighted capitalist had been present at Kitty Hawk, he would have done his successors a huge favor by shooting Orville down." Mr. Buffett here refers to the fact that the industry as a whole has produced no profits for its investors. In such a difficult environment, it is remarkable that Southwest airlines has consistently been profitable. In the 20 years ending in 2009, Southwest has generated consistent profits each year ranging from a high of $645 million in 2000 to a low of $99 million in 2009. Southwest targeted passengers that valued low fares but wanted consistent service that was friendly and punctual. To best deliver this value proposition, Southwest designed its processes to serve short haul city pairs, provide single class air transportation using only one airplane type. Planes were turned around quickly at gates to achieve much higher flying times per day than the industry average. Despite rising wages of its employees, Southwest has succeeded in maintaining high labor productivity, thus keeping its overall costs low. The alignment between its business processes and strategic position has allowed Southwest to deliver consistently positive financial results. Many other low-cost carriers have sprung up all over the world, though none has been quite as successful as Southwest. One of the more successful low-cost carriers outside the United States has been Ryanair.

The success of Southwest contrasts with the efforts of traditional carriers such as Delta and United. United entered bankruptcy between 2003 and 2006 in order to restructure and try to return to profitability. Delta lost almost $9 million in 2008. Both carriers have also attempted to set up low-cost subsidiaries with limited success. Delta set up Delta Express in

1996 to compete with low-cost carriers. It ceased operations in 2003 after Delta established Song another low-cost subsidiary. In May 2006, Song ceased operating as an independent brand and was brought back into the Delta network. In 2003, United set up its low-cost unit Ted. This effort was also short-lived with the Ted brand and services discontinued at the end of 2008.

The success of Southwest Airlines and the contrasting difficulties of traditional carriers raise a set of questions linking a firm's strategy to its business processes that this chapter attempts to address. How does a company's strategic positioning in the market affect its choice of business processes? How can a company verify that its processes have the appropriate competencies to support its competitive strategy? How can a company use the trade-offs inherent in process competencies to its advantage when designing its business processes?

The future for Southwest is likely to be somewhat more challenging than the past. To begin with, their labor costs have continued to increase while several of their competitors have managed to reduce labor costs after entering bankruptcy. Some of the new low-cost carriers (such as Jet Blue) have lower labor costs than Southwest. As Southwest has grown, the number of connecting passengers has also increased. While significantly fewer than the passengers traveling by the traditional hub and spoke carriers, about a third of Southwest passengers took connecting flights in 2009. Southwest is also considering longer haul flights, including those to vacation destinations in Mexico and the Caribbean, which will require the company to bring in new airplane types. As the Southwest example illustrates, the strategic position of a firm cannot be static and must adjust to the needs of customers being served. In order to succeed, it is, therefore, important for organizations to regularly review the design of their business processes to ensure that the appropriate competencies to support the new strategic position are being maintained and developed.

In this chapter, we examine the relationship between a firm's strategy and the design and management of its operations. Section 2.1 starts with a discussion of the concepts of strategic positioning and operational effectiveness. Section 2.2 then defines business strategy in terms of markets, customers, and products and operations strategy as planning the process architecture necessary to produce and deliver those products. In Section 2.3, we emphasize the importance of strategic fit among three pivotal aspects of a firm's operations:

1. Business strategy
2. Operations strategy
3. Process architecture and operating policies

In Section 2.4, we show how focusing on narrow market segments and structuring business processes accordingly facilitate this strategic fit. Section 2.5 presents an important example of matching products and processes according to variety and volume. In Section 2.6, we discuss the concept of operations frontier in the competitive product space of product cost, response time, variety, and quality. Section 2.7 traces the historical evolution of operations strategy and process improvements. Section 2.8 concludes the chapter by discussing some opportunities for improving service operations using the Internet and telecommunications technology.

2.1 STRATEGIC POSITIONING AND OPERATIONAL EFFECTIVENESS

The word strategy derives from the Greek military term *stratégia*—"the general's art." It is the art or science of planning a war, and much of the original management thinking on strategy treated business as something of a war and the goal was to win. However, times have changed. Chief executive officers are no longer generals, and workers are not soldiers. Today, strategy is a plan to achieve an objective (Hindle, 1994).

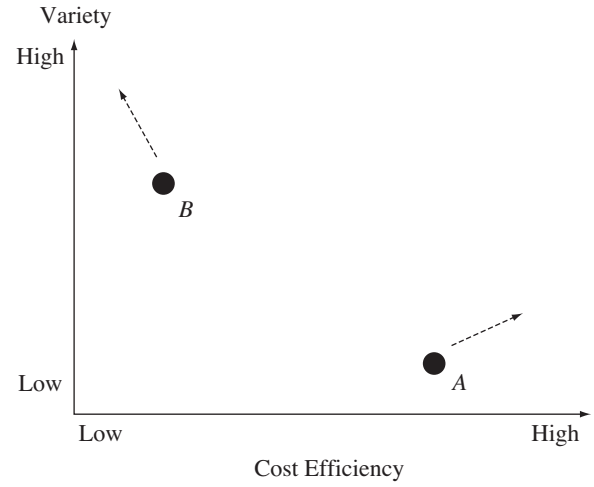


FIGURE 2.1 Current Position and Strategic Directions of Movement in the Competitive Product Space

The *plan* specifies precisely what managers must do in order to reach corporate objectives. Often, the implicit objective of a business strategy is to deliver sustained *superior*, not just average, performance relative to the competition. To outperform one's rivals, one must be—and remain—different from and better than them. Because similar firms, especially those in the same industry, perform in much the same way, a sustainable competitive advantage requires some form of differentiation.

Competitive Product Space **Competitive product space** is *a representation of the firm's product portfolio as measured along the four dimensions or product attributes—product cost, response time, variety, and quality*—that were introduced in Chapter 1. Figure 2.1 represents a firm's product portfolio in the competitive product space, but for graphical simplicity, we show only variety and cost while holding response time and quality constant. (In fact, instead of representing product cost directly, the figure shows the reciprocal of cost [1/cost] as a proxy of cost efficiency.) An organization may, for example, differentiate itself by offering customers value through a product with a unique combination of the four product attributes. Measuring and quantifying the portfolio of current product offerings along these four dimensions yields a set of *points*, one per product, in the competitive product space.

Strategic Positioning **Strategic positioning** *defines those positions that the firm wants to occupy in its competitive product space;* it identifies the product attributes that the firm wants to provide to its customers. Figure 2.1 depicts strategic positioning of two firms—A and B. Firm A provides a low-cost standardized product, whereas Firm B provides a customized but expensive product. The arrow shows the intended direction of movement as the firm's strategy.

Competitors also occupy positions in the competitive product space. One could conceivably measure product performance of each competitor, deduce its strategic positioning from the attributes of its products, and represent its current position in the competitive space. Occupying a differentiated position, then, entails producing and delivering different product attributes. This approach requires the firm's business processes to be structured and operated in ways that differ from those of competitors. In the automotive industry, for example, Hyundai aims to occupy a low-cost position, while Rolls-Royce strives for the highest-quality cars. As we will see, each company's business processes will also differ. To sustain its competitive advantage, a firm must ensure that its competition finds it difficult to imitate its chosen position.

Operational Effectiveness To deliver superior performance, a firm must strive to select product attributes that are distinct from those of its competition and create business processes that are more effective in producing and delivering them than its competition. **Operational effectiveness** means *possessing process competencies that support the given strategic position*. Developing process competencies requires designing suitable business processes and operating policies. "Operational effectiveness includes but is not limited to efficiency. It refers to any number of practices that allow a company to better utilize its inputs by, for example, reducing defects in products or developing products faster" (Porter, 1996). It is important to understand that operational effectiveness does not necessarily mean the lowest-cost process, which may be called operational efficiency. A firm such as FedEx has a strategic position and process competencies that are focused on speed and reliability, not on low cost. In contrast, Southwest has a strategic position and process competencies with a much greater emphasis on low cost. In practice, gaining and sustaining a competitive advantage requires that a firm have a good strategic position *and* operational effectiveness to support that position.

2.2 THE STRATEGY HIERARCHY

Strategy spans different levels in an organization. At the highest level of a diversified company, **corporate strategy** *defines businesses in which the corporation will participate and specifies how key corporate resources will be acquired and allocated to each business*. Corporate strategy formation is thus like portfolio selection—choosing a mix of divisions or product lines so as to ensure synergy and competitive advantage.

At the next level, **business strategy** *defines the scope of each division or business unit in terms of the attributes of the products that it will offer and the market segments that it will serve*. Here, strategy includes what we described earlier as strategic positioning. Since the goal is to differentiate the firm from its competition by establishing competitive priorities in terms of the four product attributes, business strategy entails a two-pronged analysis:

1. Competitive analysis of the industry in which the business unit will compete
2. Critical analysis of the unit's competitive skills and resources

At the next level, we have **functional strategies** that *define the purpose for marketing, operations, and finance—the three main functions in most organizations*:

- Marketing identifies and targets customers that the business unit wants to serve, the products that it must supply in order to meet customer needs, and the competition that it will face in the marketplace.
- Operations designs, plans, and manages processes through which the business unit supplies customers with desired products.
- Finance acquires and allocates the resources needed to operate a unit's business processes.

Each of these functions must translate the midlevel business strategy into its own functional requirements by specifying what it must do well in order to support the higher-level strategy.

In particular, **operations strategy** *configures and develops business processes that best enable a firm to produce and deliver the products specified by the business strategy*. This task includes selecting activities and resources and combining them into a network architecture that, as we saw in Chapter 1, defines the key elements of a process, such as inputs and outputs, flow units, and information structure. Operations is also responsible for developing or acquiring the necessary process competencies—process cost, flow time, flexibility, and quality—to support the firm's business strategy. Whereas business strategy involves choosing product attributes on which to compete,

operations strategy focuses on the process competencies required to produce and deliver those product attributes.

Thus, business strategy is concerned with selecting external markets and products to supply them, whereas operations strategy involves designing internal processes and interfaces between the input and output markets. An operations strategy must establish operational objectives that are consistent with overall business goals and develop processes that will accomplish them. For example, a business strategy based on product cost as a top competitive priority calls for an operations strategy that focuses on efficient and lean business processes. Southwest's business strategy has historically focused on low cost. To support this strategy it has designed and operated business processes that aim for high utilization of assets and a high level of labor productivity to lower costs. Similarly, if a firm seeks competitive advantage through product variety, its business processes must be flexible enough to produce and deliver customized products. For example, Zara, the Spanish apparel retailer has achieved tremendous success by providing a wide variety of products using processes that are fast and flexible enough to bring new products to market quickly and replenish them in small lots. If the goal is to provide short response times, processes must include greater investment in inventories (for manufactured goods) or greater resource availability through excess capacity (for both manufacturing and service operations) as we will show in the remainder of this book. Finally, a strategy that calls for producing and delivering high-quality products requires high-quality processes with precision equipment and highly trained workers. In every case, process competencies must be *aligned* with desired product attributes—operations strategy must be *consistent* with business strategy. Example 2.1 describes how Walmart achieved such consistency.

EXAMPLE 2.1

As an example of consistency in strategic hierarchy, consider the case of Walmart, the well-known retailer. Figure 2.2 shows how Walmart has positioned itself as a low-cost retailer of medium-quality goods supplied with high accessibility and availability in terms of both store locations and continuous product availability on store shelves. To support this business strategy, Walmart's operations strategy calls for an efficient distribution process that features short response times and low inventory levels.

> To accomplish both of these seemingly contradictory objectives, Walmart's logistics process calls for its own transportation fleet and information network, complete with satellite communications systems to connect stores in well-chosen locations. To ensure close communication among retail outlets and suppliers—and thus quick replenishment of depleted stocks—point-of-sales (*POS*) data are transmitted by a proprietary information system called Retail Link. Low pipeline-inventory levels are achieved by a system called cross-docking: incoming trucks dock opposite outgoing trucks so that goods can be transferred directly from incoming to outgoing trucks without intermediate storage.

The overall result is impressive, even when compared with other industry leaders: a high inventory turnover rate (Walmart achieved 9.2 turns in 2009 compared to 6.1 for Target), improved targeting of products to markets (resulting in fewer stockouts and markdowns), significantly higher sales per square foot of store space (Walmart averaged sales of $425 per square foot in 2009 compared to $273 for Target), dominant market

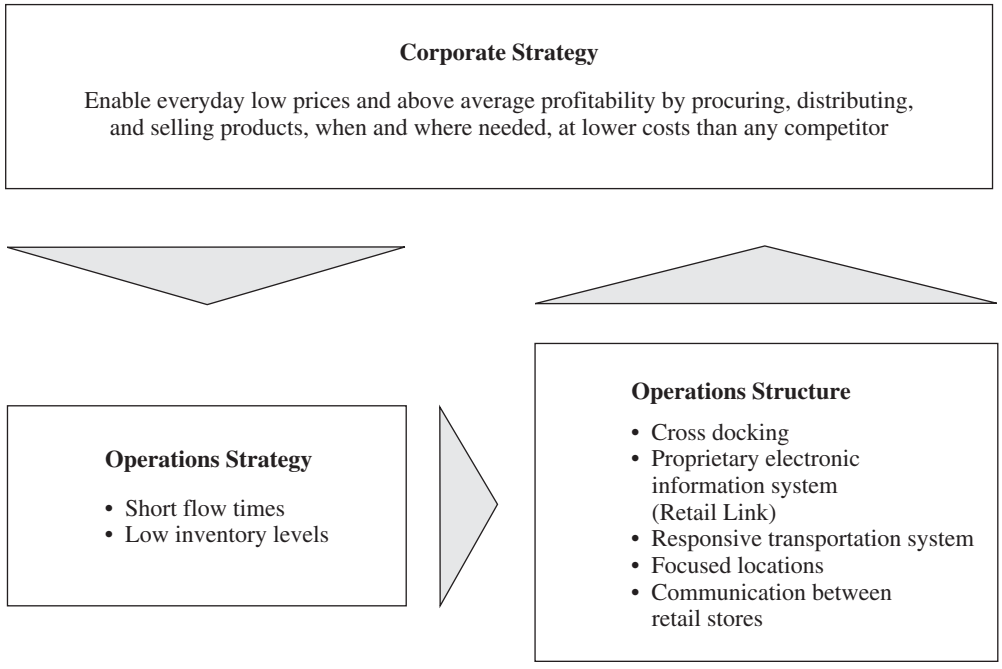


FIGURE 2.2 The Wal-Mart Strategy and Operations Structure

share, and growth (Walmart's sales in 2009 were about $405 billion compared to about $63 billion for Target). Walmart is, therefore, an outstanding example of a strategically well-positioned firm that has carefully orchestrated its operations strategy and process architecture to support its business strategy.

As they move forward, however, Walmart faces some challenges. Further growth in the United States requires Walmart to focus on smaller formats in urban areas. This is very different from the current retail network of the company that primarily consists of very large stores outside major urban areas. Walmart's current design of its business processes is unlikely to be completely consistent with the strategic position of smaller formats. Walmart will thus have to design new business processes.

2.3 STRATEGIC FIT

The hierarchical framework described in the previous section reflects a top-down approach to strategy formulation: Once the firm's business strategy has defined its position in the competitive space (as defined by price, time, variety, and quality), its business processes are then designed and managed to attain and maintain that position. It is worth pursuing this point because it helps us answer a fundamental question: What distinguishes an effective business process? In manufacturing, a common tendency is to equate an effective process with an efficient process. Although **cost efficiency**—*achieving a desired level of outputs with a minimal level of inputs and resources*—is obviously an important competitive advantage, firms may also compete on a number of other dimensions such as response time, product variety, or quality. Thus, a business process that is effective for one company may be a poor choice for another company pursuing a different strategy in the same industry.

How, then, does "effective" differ from "efficient"? A process is efficient if it operates at low cost. A process is effective if it supports the execution of the company's strategy.

A low-cost process can be both efficient and effective if, as in the case of Walmart, low cost is a key component of the strategic position of the firm. Thus, the key condition for process effectiveness is the existence of a strategic fit among three main components of a firm's strategy:

- Its strategic position
- Its process architecture
- Its managerial policies

Strategic fit means *consistency between the strategic position that a firm seeks and the competencies of its process architecture and managerial policies*. Consistency may be absent if top-level managers lack knowledge about basic business processes or if they delegate important process decisions to operating managers who are unfamiliar with the firm's overall strategy. In either case, the company's strategic position and network of business processes may be incompatible. For instance, Jaikumar (1986) gives examples of firms that had invested in flexible manufacturing systems but were still producing only a handful of products in fairly large volumes. Flexible manufacturing systems should be used to support a strategy of greater variety of products at lower volumes. Otherwise, they would simply result in an increased product cost.

The potential conflict between the top-down strategy and the principle of strategic fit was first identified in 1969 by Skinner, who argued that "too often top management overlooks manufacturing's potential to strengthen or weaken a company's competitive ability." As a result, concluded Skinner, "manufacturing becomes the missing link in corporate strategy" (Skinner, 1969). Among other things, Skinner was criticizing the perception of operations as a technical discipline concerned only with cost reduction and low-level day-to-day decisions.

Even though that misperception is still fairly widespread, consultants, educators, and practicing operations managers have made substantial progress in understanding the strategic importance of operations. Indeed, the business process reengineering movement of the early 1990s stressed the fundamental rethinking and redesign of business processes as a means of improving performance in such areas as time, cost, quality, and service. This theory advocates radical changes in processes (and, in fact, in the organization as a whole) as an effective means of formulating strategy and designing processes that will result in significant improvements in performance. By equating organizations with processes, this view has put business process design and management on the strategic agenda of top management at numerous firms (Harrison & Loch, 1995).

It is important to understand that there is no permanent state of strategic fit. Dell is a perfect illustration of the need to constantly adapt both the strategic position and the process architecture. Dell, founded in early 1984, was the worldwide leader in the computer industry with a global market share nearing 18 percent in 2004. In terms of the product attributes discussed in Chapter 1, Dell's initial focus was to increase product variety and customization while keeping product cost low and delivery-response time and quality acceptable. To best deliver that specific value proposition, Dell designed an operational process that involved direct sales coupled with a lean and responsive assemble-to-order system. According to Carpenter (2003), Michael Dell explains that "his key to success was putting the focus on the customer and building a custom computer that was exactly what the user needed." The perfect fit between intended strategic positioning and the process used to deliver the products yielded impressive returns: "[Michael] Dell said his business grew by 80 percent for the first eight years, 60 percent for the next six and about 20 percent each year since then." After ten spectacular years, Dell hit a rough patch between 2005 and 2010. Revenues increased marginally from $49 billion in 2004 to $53 billion in 2009. Annual net income, however, declined from over $3 billion in 2004 to under $1.5 billion in 2009. In fact,

Michael Dell returned to the company in 2007 to alter the two key process architecture choices that had led to success earlier. He introduced selling computers through retail stores like Walmart (instead of only selling direct) and outsourced some assembly to third parties who often built computers to stock rather than to order. These changes in process architecture were required because hardware became more of a commodity over time, and customer priorities shifted from variety (customization) to low cost. This required Dell to design new processes focused on low cost rather than flexibility.

Market- and Process-Driven Strategies Although the top-down view is convenient for explaining the concept of strategic fit, some experts urge that the relationship be reversed. Management, contends one team of researchers, should emphasize that "the building blocks of corporate strategy are not products and markets but business processes. Competitive success then depends on transforming a company's key processes into strategic competencies that consistently provide superior value to the customer" (Stalk et al., 1992).

Strategic fit may be achieved using either of two approaches:

1. **Market-driven strategy**: *A firm starts with key competitive priorities and then develops processes to support them*.
2. **Process-driven strategy**: *A firm starts with a given set of process competencies and then identifies a market position that is best supported by those processes*.

Whereas producers of commodity products tend to be market driven, technologically innovative companies tend to drive markets. Apple has had remarkable success in this regard using both its design competency and an intimate understanding of the customer to design products like the iPod, iPhone, and iPad and content delivery services like iTunes that have led the market. eBay and Google are examples of service providers whose technological innovations drove the online auction and search markets. Facebook designed technology that has led to an explosion in online social networking. In all these examples, it is important to observe that even though their origin was a technological innovation, ultimate success depended on meeting a customer need effectively. eBay used its technology to make running an auction quicker and cheaper, Google has made search quicker and cheaper, while Facebook has made connecting with others more convenient.

In general, strategic fit requires both market- and process-driven strategies. It entails identifying external market opportunities along with developing internal process competencies until the two are mutually consistent, and it means doing so repeatedly. The resulting view of strategic fit, argues one review of the field, "inextricably links a company's internal competencies (what it does well) and its external industry environment (what the market demands and what competitors offer)" (Collis & Montgomery, 1995).

2.4 FOCUSED OPERATIONS

The concepts of strategic fit and strategic positioning are rooted in the very existence of trade-offs and the need to make choices. As discussed, strategic fit requires business processes that are consistent with a given business strategy. However, because no single process can perform well on every dimension, *there cannot be a process that fits all strategies*. Choosing a strategy, therefore, involves focus: "The essence of strategy," observes Michael Porter, "is what to do and what *not* to do" (Porter, 1996).

Focused Strategy and Focused Processes It is generally easier to design a process that achieves a limited set of objectives than one that must satisfy many diverse objectives. This fact underlies the concept of **focused strategy**: *committing to a limited, congruent set*

of objectives in terms of demand (products and markets) and supply (inputs, necessary process technologies and volumes). In other words, this approach concentrates on serving limited market segments with business processes specifically designed and operated to meet their needs.

In turn, a focused strategy is supported by a **focused process**—*one whose products fall within a small region of the competitive product space*. All products from a focused process have similar attributes in terms of cost, quality, response time, and variety. The area occupied in the product space by the product portfolio of a focused process is small. Conversely, if the product portfolio is more dispersed in the competitive space, then the process is less focused. The Aravind Eye Hospital in Madurai, India, offers a good example of a service operation that is focused on providing high quality and low price at the expense of variety. In 2009, Aravind provided over 300,000 cataract operations. According to Rubin (2001), its founder, Dr. Govindappa Venkataswamy (also known as Dr. V.) specialized his surgical instruments and his "process of cataract surgery" which allowed him to do as many as 100 surgeries a day! Aravind's operational excellence yields gross margins of 40 percent despite the fact that 70 percent of its patients pay nothing or almost nothing and that the hospital does not depend on donations.

An example of a ferocious cost competitor is Aldi, an international retailer specializing in a limited assortment of private label, high-quality products at the lowest possible prices. The first Aldi store opened in 1948 in the German town of Essen. Today, Aldi is a leader in the international grocery retailing industry with more than 8,000 stores and has operations in Europe, the United States, and Australia. Its secret to success is found in a lean operating structure with emphasis on frugality and simplicity that yields a very low-cost structure. Brandes (1998) describes Aldi's founder Theo Albrecht as "a man who uses paper on both sides and likes to turn off the lights when leaving a room." Low prices, however, come at the expense of much smaller variety and assortment and lower availability (in terms of frequent stockouts) than competitors.

While offering a narrow product line is the most intuitive example of focus, a focused process need not just produce a limited number of products. A job shop (discussed in Chapter 1) can be viewed as a focused operation, whose mission is to provide a variety of products as long as they all have similar quality, cost, and timeliness attributes so that they all fall in a small area in the product space. Similarly, emergency rooms in hospitals focus on variety and responsiveness in the service sector. In July 2000, the emergency department at Oakwood Hospital in Dearborn, Michigan (www.oakwood.org), began guaranteeing that an emergency room patient will see a physician and have treatment started within 30 minutes. Oakwood completely revamped emergency room processes. The result was that the average wait to see a physician in the emergency room shrank from several hours to 22 minutes, and a year later, all but 32 out of 60,000 emergency room patients saw a physician within the guaranteed 30 minutes. Another example of a firm focused on providing speed and variety but not low cost is materials, repair, and operations products distributor McMaster-Carr, which we discussed in Chapter 1.

Even if a strategy calls for serving broad market segments, each of which requires a different strategic emphasis (or positions in the competitive product space), it can be separated into substrategies, each focusing on a limited, consistent, and clear set of objectives. Each substrategy can then be supported by its own consistent—or focused—business process. Depending on the scale of the business, this approach leads either to a focused plant that performs one specific process or to a **plant-within-a-plant (PWP)**, *in which the entire facility is divided into several "miniplants," each devoted to its own specific mission with a process that focuses strictly on accomplishing that mission*.

Most general hospitals, for example, have realized the benefits of focus by separating the emergency room and trauma units from the rest of the facility. Some hospitals

(such as Massachusetts General) have gone even further by developing separate units to focus on such routine operations as hip or knee replacement and rehabilitation. Many hospitals have set up separate ambulatory cancer care units that are able to provide ambulatory patients with a very high level of service at relatively low cost. Separating simpler ambulatory cases has also improved the responsiveness that hospitals are able to offer to the truly complex cases. Specialty hospitals, such as Shouldice Hospital that we mentioned in Chapter 1, which provides only hernia repairs, have only one such focused unit.

Many manufacturers also choose to separate product lines within a plant. Harley-Davidson maintains two separate flow processes: one for its small 833cc engine and transmission systems and one for its large 1,340cc power trains. This strategy is logical because each product line requires a different process and has sufficiently high volume to warrant the investment. Similarly, engine maker Briggs & Stratton separates its various production plants and, within each plant, separates product assembly lines in PWPs.

Finally, achieving strategic fit through focused operations provides firms with a powerful deterrent barrier against competitors' efforts to imitate them. Their competitive advantage, therefore, is more sustainable. Although any single activity may be vulnerable to imitation, the greater the number of activities involved, the harder the wholesale imitation becomes. Supporting a firm's strategic position with multiple, mutually reinforcing activities creates sustainable competitive advantage because it is, according to Porter, "harder for a rival to match an array of interlocked activities than it is merely to imitate a particular [activity]" (Porter, 1996). Indeed, copying a focused business process—a complete network of activities, resources, and managerial infrastructure—amounts to cloning the entire organization. If a firm's process and strategy are both focused, its position is already the result of carefully considered trade-offs made when managers chose their position and its supporting process. (We will discuss more about trade-offs in Section 2.6.) A competitor, therefore, can copy that position only by making similar trade-offs. In so doing, it will inevitably be giving up its own position. Example 2.2 describes United Airlines' unsuccessful attempt to imitate its low-cost-focused competitor, the Southwest Airlines.

EXAMPLE 2.2

In 2004, United Airlines created Ted, an "airline within an airline" to compete with low-cost carriers like Southwest Airlines. Ted was equipped with 57 Airbus A320 aircraft in an all economy configuration. All Ted flights, however, were operated by United crew. Equipment substitutions often led to United aircraft being operated as Ted flights. Customers were also confused by the relationship and often connected between Ted and United flights. The crossover of passengers, crew and aircraft made it difficult for Ted to be a truly focused "airline within an airline." While United was able to copy some aspects of the processes that made low-cost carriers like Southwest successful, it could not replicate them all because of the absence of focus. The result was that United could neither satisfy business travelers who ended up on Ted expecting it to be like United nor lower costs to the level of focused low-cost carriers. United eventually announced the dismantling of Ted in June 2008.

As illustrated by their failure in setting up sustainable low-cost operations, traditional airlines have had difficulty implementing the PWP concept. This is largely because of their inability to truly separate the low-cost processes from the processes focused on the traditional customer. Whereas Southwest was able to run a point-to-point network, the traditional carriers had to allow customers to connect between their low-cost routes and other flights. Not only were passengers exchanged, but also the

crew and planes often moved between processes. This made it very difficult to lower labor requirements necessary for the low-cost process because flights connecting at a hub result in a very uneven workload if connection times are to be kept short. Also, delays tended to propagate across the entire network unlike for Southwest where limited connections allowed the impact of delays to be isolated.

Today it may be argued that Southwest is facing some of the same challenges as it has grown larger. Whereas the company historically had very few connecting passengers, by 2009 about a third of passengers were connecting to other flights. While this was still about half the number for traditional airlines, it does impose challenges to running a pure point-to-point airline.

2.5 MATCHING PRODUCTS AND PROCESSES

Focused operations make it easier for a firm to match its processes with the products that it produces. *A useful tool for matching processes to products* is the **product–process matrix** proposed by Hayes and Wheelwright (1979). A model of this matrix is shown in Figure 2.3.

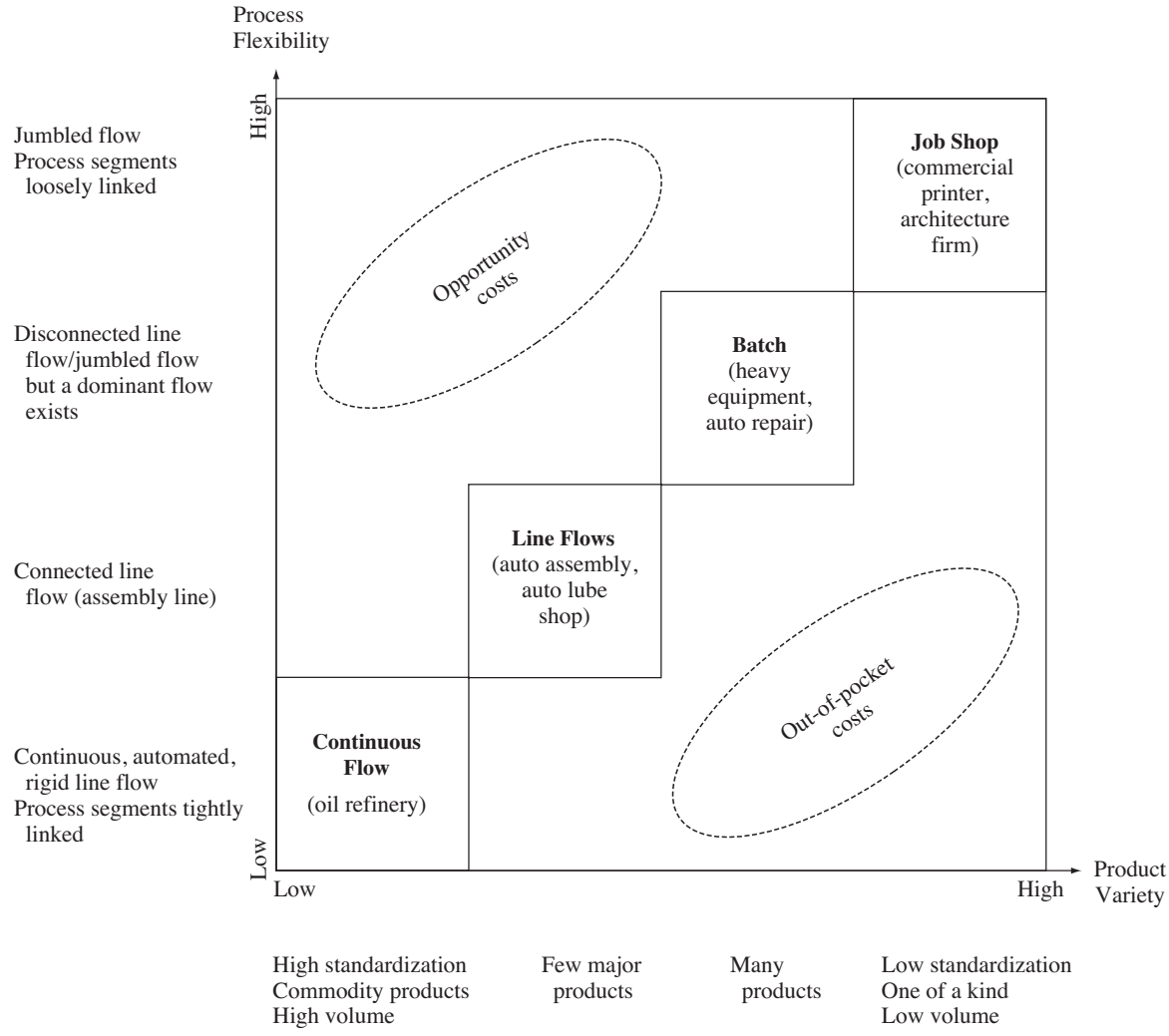


FIGURE 2.3 The Product-Process Matrix

The horizontal axis charts product variety from "low variety" (representing standardized products produced at high volume) to "high variety" (representing one-of-a-kind products produced at low volumes). At the right end, we would find such unique products as skyscrapers, tailor-made suits, consulting reports, and plastic-surgeries. Such highly customized products are demanded and produced to order, one at a time. At the left end is the other extreme: highly standardized commodity products demanded and produced in bulk. Beer breweries and commercial paper mills illustrate this end of the spectrum. Between these extremes fall products that have intermediate degrees of customization and volume. Thus, houses built by a real estate developer may be largely similar to a limited variety of model homes while permitting some degree of customization and upgrades.

The vertical axis measures process flexibility, the process competency on the supply side that corresponds to product variety. At the bottom, low process flexibility results from a process architecture with rigid, fixed routes and specialized resources that can perform only a narrow set of tasks repeatedly, as in a flow shop. At the top, high flexibility results from processes that employ general-purpose resources that are loosely linked so that products can flow along many different routes, yielding "a jumbled work flow," as in a job shop (see Chapter 1). Intermediate processes differ in the degree of specialization, volume capability, and interlinkage of resources.

Ideally, each process type fits a specific product demand: Job shops, for instance, are ideally suited to produce custom products in low volumes, while flow shops work best for more standardized products demanded in high volumes. Effective product–process matches occur on the diagonal of the product–process matrix. An off-diagonal position represents a mismatch that can result in unnecessarily high costs. Thus, a flexible job shop that produces only one product results in opportunity costs of not producing a wider variety. Similarly, a specialized flow shop that produces several products in low volumes undergoes numerous equipment changeovers, resulting in out-of-pocket costs. A diagonal position corresponds to a proper match between the desired product variety and the necessary process flexibility.

Note that the product–process matrix connects only one product attribute with one process competency. There is also a correlation between process flexibility and product cost: standardization typically results in economies of scale and thus lower variable product cost. Likewise, there is a correlation between process flexibility and product response time: flow shops typically have shorter flow times than job shops. Product quality, however, bears no direct correlation to layout of resources and connecting routes. Both job shops and flow shops can produce high quality.

2.6 THE OPERATIONS FRONTIER AND TRADE-OFFS

Once the firm has chosen its operations strategy and process architecture, it must operate the process to execute the strategy. As discussed, a strategic position supported by consistent business processes that are managed effectively is essential for superior performance. Sustained competitive advantage requires *both* good strategic positioning and operational effectiveness. Strategic positioning is about choosing a different set of customer needs to serve or choosing to serve existing needs in a more effective manner. Firms change strategic positions infrequently. When managers are considering such a change, they ask, "*What* should we do and not do?" Operational effectiveness, on the other hand, is about structuring processes to best support the chosen strategic position and then executing these processes better than rivals. When managers are considering changes to the operating policies of a process structure already in place, they ask, "*How* could we better design and manage our business processes?"

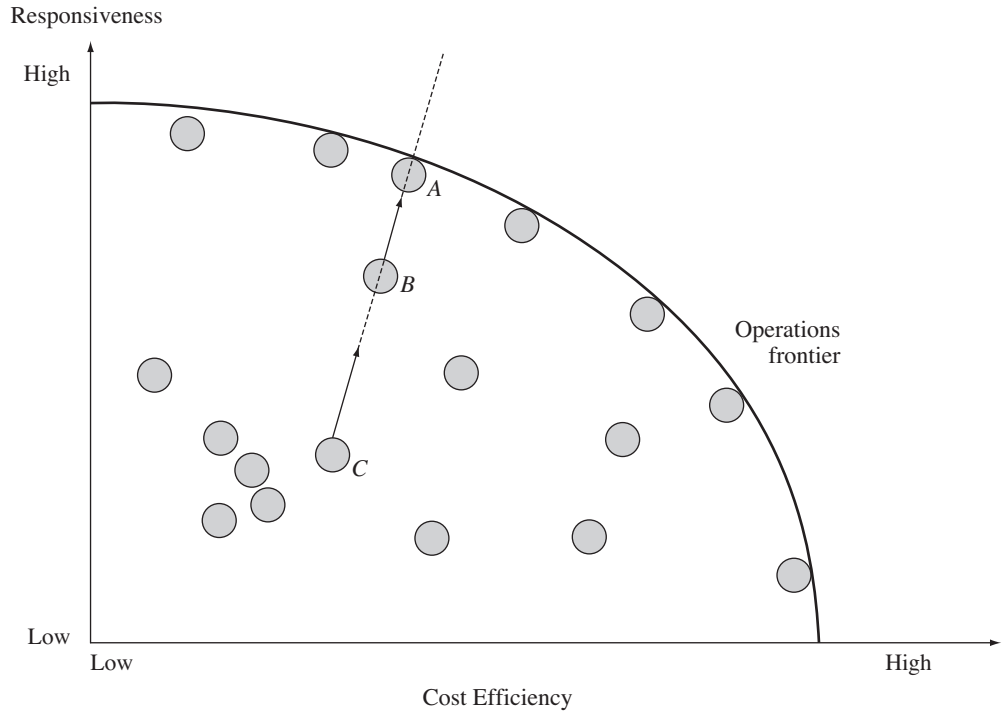


FIGURE 2.4 The Operations Frontier as the Minimal Curve Containing All Current Positions in an Industry

The Operations Frontier Earlier, we represented a strategic position by the location of the firm's products in the competitive product space. An empirical study of a particular industry might measure and position each firm's product offerings in that space. In Figure 2.4, we illustrate a variety of product offerings using the two dimensions of responsiveness and cost effectiveness. In general, such a picture can be visualized with all four dimensions of cost, quality, response time, and variety. One could then define the **operations frontier** as *the smallest curve that contains all current industry positions*. It represents the current best practices of world-class firms. Firms located on the same ray share the same strategic priorities. However, firms operating on the operations frontier boast of superior performance: They have the highest operational effectiveness—the measure of how well a firm manages its processes. Their processes provide superior performance along the desired product attributes. Operational effectiveness is thus related to the distance of the current position from the (current) operations frontier. The closer a firm is to the frontier, measured along its direction of improvement (whose slope represents the relative strategic priorities assigned by the firm to the four dimensions), the higher its operational effectiveness. In Figure 2.4, the companies delivering products A, B, and C share strategic priorities, yet the company delivering product A has the highest level of operational effectiveness. It defines the current best practices to manage its business processes (since it is on the frontier), while the product C company has the lowest level of operational effectiveness (as it is farthest from the frontier, as measured along its direction of movement).

Trade-Offs A **trade-off** is *a decreasing of one aspect to increase another*. Because the operations frontier is typically concave, any point on the frontier represents a trade-off: To increase performance along one product dimension, one must give up some performance along the other(s). It thus follows that firms that are not on the frontier do not

face trade-offs: They can improve along multiple dimensions simultaneously. Trade-offs, therefore, are typically reflected most clearly in the strategies of world-class companies, such as Toyota, as described in Example 2.3.

EXAMPLE 2.3

In the late 1960s and early 1970s, Toyota, a small Japanese automobile maker, was facing a depressed economy in a country where space is at a premium. Because no firm can survive producing only a single product for small, depressed markets, product variety was a necessity. So Taiichi Ohno and his coworkers developed the *Toyota Production System* (*TPS*). The key idea behind TPS was to produce exactly what you need (regardless of variety) exactly when you need it. The potential problem was equally simple: There was no room for error. Suppliers and equipment had to be reliable, production had to be flexible, quality had to be high, and consistency was necessary in every respect. Critical to the success of the system were carefully coordinated interactions with suppliers, who had to meet both precise delivery schedules and precise performance specifications while remaining as flexible as the automaker itself.

TPS was, in fact, the reinvention of Henry Ford's assembly line or process-flow concept (see Section 2.7), though with an important modification: Instead of focusing on low cost and no variety, TPS allowed product variety through process flexibility. TPS simultaneously permitted wide variety, high quality, low cost, and short response time. In effect, it so completely redefined the operations frontier that competitors all over the world had to scramble to catch up. Having established TPS as the world-class flow process for discrete manufacturing, Toyota remains an excellent example of a company that used manufacturing as a competitive weapon in rising from obscurity to the top ranks of its industry.

Initially, when competitors began copying elements of TPS, they saw the possibility of dramatic improvements in both cost and quality. They thought that if such operational effectiveness was possible, perhaps the traditional trade-offs between cost and quality or cost and variety were no longer valid. Since most of these rivals were far from the best in their class, they originally did not have to make genuine trade-offs in their quest for operational effectiveness. Their operations were so ineffective that they could simultaneously improve on several dimensions of their processes.

Improved operational effectiveness is not the same as improved strategic positioning. Whereas strategic positioning defines the direction of improvement from the current position, improving operational effectiveness reduces the distance of the current position to the current operations frontier along the direction of improvement. When a firm's position on the operations frontier is developed according to the "state of best practices," it represents the best attainable trade-off between the two dimensions at a given point in time. Example 2.4 describes the trade-offs that differentiate an emergency room from other hospital processes.

EXAMPLE 2.4

If a general hospital tries to handle emergency and nonemergency cases with a single process, its products will have very different strategic emphases. Consequently, such a process will cover too large an area in the competitive product space and make it difficult for the hospital to be competitive on all dimensions. Suppose, however, that the

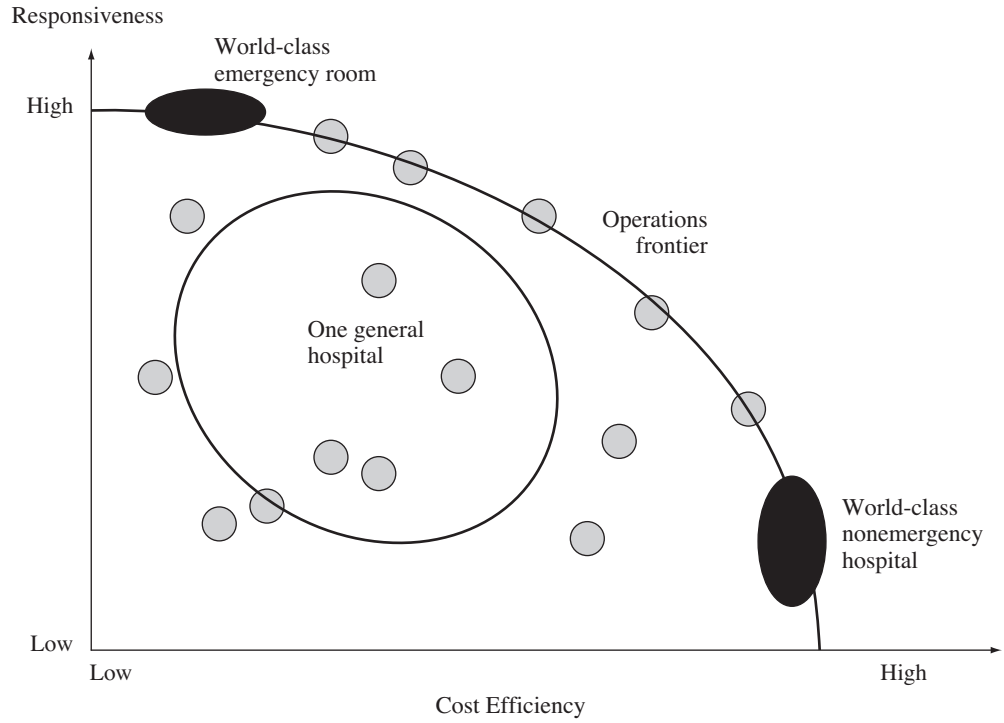


FIGURE 2.5 The Operations Frontier in the Health Care Sector

hospital divides its operations into two distinct plants-within-a-plant (PWPs)—emergency-room and nonemergency facilities. Suppose, too, that each PWP has its own competitive priorities and a consistent process to support those priorities. Clearly, an emergency room will employ doctors and staff who are "on call" and will have the flexibility to treat a wide variety of cases rapidly. A general hospital, meanwhile, can afford more specialized doctors, each geared to treating a small set of cases. In Figure 2.5, the products of each PWP now share similar competitive priorities and thus occupy a smaller area. Each PWP process is more focused, and it is easier for each process to perform effectively its particular strategic mission.

Improvements in operational effectiveness bring a company closer to the frontier or move the frontier itself along the direction of improvement specified by the strategic position. By including direction as part of operational effectiveness, we measure alignment between strategy and process competencies. As such, operational effectiveness is at the core of superior performance. It is important to maintain alignment as we improve processes. This is illustrated by Toyota's Global Body Line (GBL), implemented in 2002 to allow greater variety at low cost (Visnic, 2002). GBL has a stated goal of allowing Toyota to manufacture their products in any country in any volume. While speeding up the flow of a vehicle through the body shop, GBL has also significantly reduced the time required to switch models or complete a major model change. GBL has moved the operations frontier by improving both the efficiency as well as the flexibility of its process.

As a company improves its processes along certain dimensions, it is important to ensure that operational effectiveness is maintained and performance does not suffer along any dimension important to customers. Toyota itself faced significant challenges around 2009 as it had to recall about 12 million cars worldwide because of quality problems. Common parts used in these cars helped handle variety at lower cost but also

increased the negative impact of one of the common parts having a quality problem. The price paid by Toyota of having a quality recall that included North America, Europe, and Asia was very significant.

As technology and management practices advance, the operations frontier shifts outward, or away from the point of origin in the competitive space. World-class companies achieve either the same response time (or quality or variety) at a lower cost or better response time (or quality or variety) at the same cost. As the dynamics of competition keeps pushing the operations frontier outward, world-class firms must continuously improve operational effectiveness merely to maintain their current positions.

The Internet is an example of a new technology that has shifted the operations frontier outward. As Chopra and Van Mieghem (2000) point out, however, the value gained by adopting electronic commerce has varied by industry. For example, the advent of Internet grocers such as Peapod in the United States and Tesco in the United Kingdom clearly enhanced quality of service to customers, but the convenience of home-delivery service typically comes at an increased cost and reduced responsiveness and variety compared to a regular supermarket store. In 2003, Peapod offered about 10,000 items, whereas a regular U.S. supermarket carried about 40,000 items. Several grocery delivery businesses such as Webvan that tried to compete with supermarkets on price have failed. Thus, while the adoption of the Internet in the grocery home-delivery business increased quality of service to consumers, it also increased cost and reduced responsiveness and variety. In 2007, the Amazon started AmazonFresh to deliver groceries in the Seattle area. While Amazon will be able to compete effectively with supermarkets in terms of the convenience it offers customers, it will find it much more challenging to serve cost conscious customers nationwide given the high transportation cost of home delivery.

In the book industry, however, the impact of the Internet is quite different, as illustrated by Example 2.5, which compares a regular Walmart Supercenter store with its Internet store. In the publishing industry, the adoption of Internet technology increases service and selection, and thus pushes the frontier out along the dimensions of service and variety. In fact, with content becoming digital, it is possible today to download an e-book in minutes at a cost that is significantly lower than the cost of printing and distributing a traditional book.

Another example where the Internet has fundamentally altered business processes is the rental and sale of movies. Traditionally, large video rental stores such as Blockbuster were the destination of choice when people wanted to rent a DVD. By 2010, there were several different options facilitated by the Internet that brought down Blockbuster. Through Netflix, one could order DVDs on the Internet that were delivered by mail. While customers had to wait for the DVD, Netflix offered significantly higher variety than Blockbuster. For some movies, Netflix also offered streaming that allowed customers to instantly watch any available movie. Another option to Blockbuster was Redbox where movies were rented through vending machines typically located at grocery stores or fast food restaurants (there were far more Redbox vending machines than Blockbuster stores allowing customers to find one nearby). Redbox offered limited variety but customers could go online and identify a vending machine nearby that had the movie they wanted available. The movie could be reserved online and rented for a dollar a day. Both Netflix and Redbox have introduced business processes that have moved the efficient frontier relative to Blockbuster's position by allowing improved responsiveness at a lower cost.

Finally, in many business-to-business settings, the Internet allows improved responsiveness and accuracy of information exchange, which can translate into both cost savings and faster order fulfillment. Collaborative planning, forecasting and replenishment between Walmart and Proctor & Gamble through information (demand,

inventory, and capacity) exchange on the Internet has allowed both companies to better match supply and demand resulting in lower costs and improved product availability.

This book is about understanding how changes in business processes result in improvements in the product value relative to the cost of production and delivery. Given that the Internet is just another channel for delivery of information, our general principles can be applied to determine the value and cost impact of electronic commerce, as the remainder of the book will illustrate.

EXAMPLE 2.5

As an example of the Internet's impact as a new channel for value creation on the operations frontier, consider Walmart.com, the Internet store of the well-known company that was introduced in Example 2.1. Walmart.com started selling merchandise online in July 1996, but replicating its off-line success was not easy. The online customer base is different from its traditional counterpart, and Internet stores require different technology than regular stores. Seeking greater online expertise, Walmart spun Walmart.com off as a separate company in January 2000.

Walmart soon became convinced, however, that its true strength online is in bricks-and-clicks integration. So in July 2001, it bought back outside stakes and turned Walmart.com once again into a wholly owned subsidiary of Walmart. Now, for example, a customer can choose replacement tires online and have them installed at a local Walmart. In 2008, about 40 percent of Walmart.com customers picked up their orders for free from a nearby Walmart store. By 2008, Walmart.com controlled nearly 8 percent of the online holiday retail traffic, behind only Amazon, which had about 15 percent.

While a typical Walmart Supercenter stocked about 100,000 items in 2003, Walmart.com stocked about 600,000. Most of that increase in variety, however, is in the 500,000 book titles and 80,000 CDs that Walmart.com carries, as compared to thousands stocked at regular stores. In addition, because of added transaction costs (order-sized pick, pack, and transportation) Walmart.com eliminated true low-cost items costing a few dollars or less. As illustrated in Figure 2.6, a single Internet channel for order taking

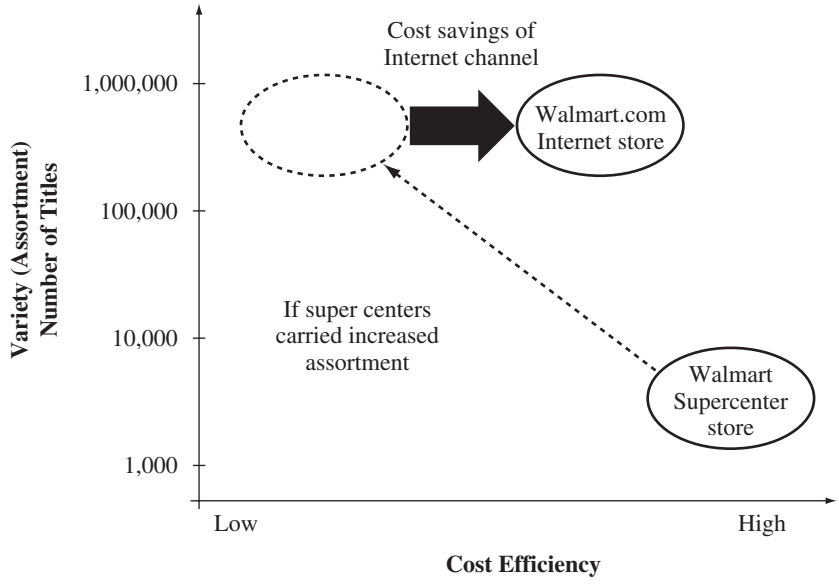


FIGURE 2.6 The Internet Channel versus Traditional Retail Stores

together with a centralized warehouse for fulfillment allows greater variety in the book and CD segment at a lower cost than physical stores could provide. Compared to the fast-moving best-sellers offered in regular stores, increasing variety naturally requires stocking more slow-moving items that have less predictable sales. The Internet channel allows Walmart.com to centralize storage and fulfillment of those slow-moving items. Chapters 6 and 7 will show that the sources of those savings can be substantial and accrue because of increased scale economies and statistical benefits of pooling on inventory. The end result is that adoption of Internet technology in this industry has the effect of pushing out the operations frontier along the variety and cost dimensions.

2.7 THE EVOLUTION OF STRATEGY AND OPERATIONS MANAGEMENT

Over time, business strategies and processes will change in response to changes in the company's industry or in its technology. In particular, the historical evolution of operations and process management is linked intimately to technological changes and their role in the development of industrialization. Until 1765, the world of commerce and technology had changed very little from the one known to the Greeks and the Romans. Although advances had been made in the textile, printing, and construction industries, the commercial world of 1765 still faced the same physical limitations as ancient cultures. Transportation speed—whether of people, of goods, or of information—was limited on land by the speed of the fastest horse and on water by the most favorable winds. At the end of the 16th century, for example, the trip from New York to Boston—a distance of 288 kilometers, or 175 miles—took three days (Ambrose, 1996).

The Factory System and Specialization This situation was destined to change dramatically in 1765, when the factory system heralded the start of the industrial revolution and the end of the "artisan" system consisting of craft guilds and decentralized cottage industries. The factory system was the result of three innovations:

1. Scottish economist Adam Smith proposed that the **division of labor** and **functional specialization**—*a process and organizational structure where people are specialized by function, meaning each individual is dedicated to a specific task*—would lead to vast improvements in cost and quality, albeit at the expense of flexibility. (The other organizational structure is **product specialization**—*wherein people are specialized by product, meaning each individual is dedicated to perform all functions on a specific product line*.)
2. Scottish engineer James Watt's invention of the steam engine made it possible for powered machinery to replace human labor. The transportation speeds of goods carriers powered by steam soon increased by a factor of 20. (Meanwhile, the telegraph removed virtually all limits on the transmission speed of information.)
3. The practice of centralizing work in one facility. This practice facilitated economies of scale and led to the growth of the assembly line and **mass production** of *large quantities of goods at low cost*.

From Standardization to Mass Production In 1810, based on innovations by Eli Whitney and Samuel Colt at the national armory at Springfield, Massachusetts, the **American system of manufacturing** *introduced the use of interchangeable parts, thereby eliminating the need to custom-fit parts during assembly*. Standardization had begun. The end of the nineteenth century brought technological advances that were prominent in such commercial phenomena as the "bicycle boom" of the 1890s—sheet-metal stamping and electrical-resistance welding allowed for both new designs and assembly methods. Another fundamental change occurred on April 1, 1913, when Henry Ford introduced the moving assembly line—the first machine-paced flow shop in manufacturing—at his

plant in Highland Park, Michigan, and thus dawned the era of mass production. "Armory practice and sheet steel work," reports one survey of U.S. business history, "equipped Ford with the ability to turn out virtually unlimited numbers of components. It remained for the assembly line to eliminate the remaining bottleneck—how to put these parts together" (Hounshell, 1984). ("Disassembly lines" had appeared earlier in the cow meat stockyards of Chicago at the end of the nineteenth century.)

Ford's primary mode of competition soon became low cost. Scale economies and the virtual elimination of all product flexibility made cars available in high volume for a mass market. Prior to the development of the assembly line, for instance, a Ford Model T required 12.5 hours of assembly-worker time—a limitation that the assembly line reduced to only 1.5 hours. Soon, Ford's plant in Rouge, Michigan, was a totally integrated facility with the best furnaces, operational equipment, and electrical systems, efficiently converting raw materials into cash in 36 hours. It was also a highly focused plant serving a competitive strategy of low cost but no variety. It produced only one product, the Model T, and Henry Ford's attitude toward the higher costs entailed by product variety was uncompromising. Of the Model T, he said, "You can have any color as long as it's black."

Flexibility and the Productivity Dilemma The changeover from Model T to Model A in 1927 was the end of Ford's competitive advantage. Alfred Sloan of General Motors had introduced the concept of "annual models" and the slogan "a car for every purpose and every price." The practice of **flexible mass production**—*a method of high-volume production that allows differences in products*—introduced product variety as a second mode of competition in the automobile industry. It was accompanied by one of the most significant trade-offs in the history of strategic positioning: Faced with the so-called **productivity dilemma**, manufacturers were obliged to *choose between the lower productivity entailed by frequent product changes or the higher productivity that was possible only if they declined to introduce variety into their product lines* (Hounshell, 1984).

From Scientific Management to Employee Involvement The first few decades of the 1900s also witnessed the rise of scientific management, which was based on the time and motion studies conducted by Frederick W. Taylor at the turn of the twentieth century. Taylor's philosophy centered on three ideas (Hounshell, 1984):

1. Scientific laws govern how much a worker can produce per day.
2. It is management's function to discover and apply these laws to productive operations systems.
3. It is the worker's function to carry out management decisions without question.

Taylor's "ceaseless quest for the 'one best way' and efficiency changed the very texture of modern manufacturing. Taylor influenced Ford's assembly line" (Kanigel, 1997) and led universities to start new "industrial engineering" departments. His ideas of industrial organization and scientific observation inspired the **statistical quality control** studies—*a management approach that relies on sampling of flow units and statistical theory to ensure the quality of the process*—of Shewhart at Bell Laboratories in the 1930s and Elton Mayo's celebrated Hawthorne studies of worker motivation at Western Electric, which highlighted the importance of employee involvement and incentive systems in increasing labor productivity.

Competitive Dimensions After World War II The period after World War II found the United States with a virtual monopoly over worldwide productivity. With most of Europe and Japan practically destroyed, there was no competition for meeting pent-up consumer demand. Thus, high demand and scale economies rose to the top of the American strategic agenda. The 1960s witnessed the rise of enormous integrated economic

structures and the emergence of huge capital investments as the main barrier to entry in many industries.

During the 1970s, Japanese manufacturers began to incorporate quality into their cost-focused strategy. Toyota began developing what, in Example 2.3, we described as the Toyota Production System (TPS). Among other things, TPS gave rise to a fourth competitive dimension: the use of time in product development, production, and distribution. The emergence of Japanese manufacturing as a global force in the 1980s led to a renewed interest in manufacturing as a competitive weapon in the rest of the industrialized world. It gave rise to a variety of new management philosophies and practices, including total quality management (TQM), just-in-time (JIT) manufacturing, time-based competition, and business process reengineering. In addition, new technologies like computer-aided design and manufacturing (CAD/CAM), flexible manufacturing systems, robotics, and Internet-based processes now play important roles in modernizing business processes.

The Growth of Information Technology The late twentieth century and the beginning of the twenty-first century witnessed the growth of **information technology**, *the hardware and software used throughout businesses processes to support data gathering, planning and operations*. In the late twentieth century, companies invested significant money and effort to implement **enterprise resource planning** (ERP) systems that *gather and monitor information regarding materials, orders, schedules, finished goods inventory, receivables, and other business processes across a firm*. As firms grew, it had become increasingly harder to coordinate the operation of different functions within a firm. The implementation of ERP systems facilitated this coordination across business processes. The accurate availability of transaction data also helped improve planning processes related to inventory availability, demand, and production across the supply chain. From demand planning supported by customer relationship management systems, to production planning supported by manufacturing execution systems, to product development and procurement supported by supplier relationship management systems, information technology has become a significant enabler of every business process within a firm.

Information technology has facilitated design collaboration and innovation efforts across supply chains. An excellent example is given by Billington and Jager (2008). Goldcorp Inc., one of the world's largest gold producers, wanted to improve productivity in their mines. They broadcast detailed information about their mines and offered a prize for the best ideas. Two Australian companies collaborated to come up with innovations that allowed Goldcorp to increase production by a factor of almost 10. Information technology allowed talent half way across the world to provide innovative ideas that improved Goldcorp's mining processes. Today it is common for multinationals to have their design and innovation efforts distributed all over the world to take advantage of global talent and ideas.

During the early twenty-first century, the growth of the Internet and cell phones transformed the interface between business processes and the customer. Business processes can instantaneously communicate with the customer at very low cost allowing for more customized communication related to products and prices. An interesting example in this context is Groupon, a Web site focused on local advertising. Historically, small local businesses have relied on print advertising to get word out about their businesses. Given the time and cost involved, success from print advertising has been somewhat limited. Groupon, founded in 2008, allows local businesses to send a "deal of the day" to the Web site's members instantaneously and at low cost. Small businesses reported significant success in reaching customers but more time is required to fully understand the profit impact of these efforts. Social networking sites like Facebook and Twitter have become an important channel through

which firms communicate with their customers instantaneously. One of the challenges posed as a result is the increased expectations from customers about the responsiveness of business processes. Given that customers can communicate with a process instantaneously, they also expect output from the business process much quicker than they did in the past. This has increased pressure on business processes to be more responsive. Firms like Amazon and Zappos have achieved considerable success by developing responsive business processes that meet customer needs quickly and accurately.

2.8 THE OPPORTUNITY TODAY IN SERVICE OPERATIONS

The beginning of the twenty-first century has seen a transformation in service processes facilitated by the growth in the Internet and telecommunications technology. The technological changes have allowed service processes to be designed and executed in a manner that provides increased access while lowering the production and delivery costs and improving the response time. The result has been an explosion in new services being offered at both the high and the low end of the economic spectrum.

One example is Ovi Life Tools offered by Nokia in countries like China, India, Indonesia, and Nigeria. In India, a farmer can use Ovi Life Tools on his cell phone to obtain weather information and wholesale prices for his produce for under $1.50 per month. The knowledge of wholesale prices allows the farmer to maximize revenues from his produce by targeting the right locations and times to sell. Nokia is also planning to offer a certificate program in English in conjunction with the Indira Gandhi National Open University in India. As these examples illustrate, improved communications have made the transfer of information goods both cheaper and quicker while increasing access. In developing markets, this technology has allowed access to poorer customers who could not be served without very low-cost processes made possible by the Internet and cell phones. In developed markets, technology has transformed the distribution of books, music and movies with physical products (books, CDs, and DVDs) being replaced by electronic goods that can be downloaded quickly at much lower cost. For example in December 2010, season one of The Monk television series DVD set was available on Amazon for $34.49 but could be downloaded for $11.88. In each of these examples, technology has significantly reduced the resources required in the transfer of information goods (thus reducing costs) while simultaneously decreasing the delivery time.

A second set of innovations has been facilitated by standardization and subsequent automation of processes delivered via the Internet. A classic example is eBay and its impact on auctions. In the first six months of 2005, eBay members in the United States sold merchandise worth approximately $10.6 billion. More recently, online auctions have also grown significantly in both China and India. Auctions were traditionally considered to be highly customized processes that could not be initiated until the seller of the good arrived. The auction was also executed with bidders or their representatives physically present in a room along with the auctioneer. eBay took the auction process apart and standardized most of the process except for the seller, the product description, and the reserve price. The standardized processes were automated on the eBay Web site allowing sellers to set up an auction on their own and bring in buyers who could be located in any part of the world. Standardization and automation significantly reduced the resources and time required to organize and execute an auction. The resulting process made it economical to buy and sell products worth as little as a few dollars. This is in contrast to traditional auctions where a large batch of products was auctioned off together to gain economies of scale and attract a large number of buyers.

Many services have become more efficient by reorganizing their processes and by using technology to allow better utilization of expensive resources. For example, McDonald's moved the order taking function at many drive though lines from inside the restaurant to a remote location. Cheaper communications allowed the company to create a centralized staff of specially trained order takers in a remote location who entered the order which showed up instantaneously on the production screen of the restaurant. Centralizing the order takers allowed McDonald's to cut labor costs and yet improve customer service, because an idle order taker could now take an order from any McDonald's restaurant. A similar idea was used by Aravind Eye Hospitals (AEH) to extend coverage in rural areas through telemedicine (see Anupindi et al., 2009). Aravind Eye Hospitals historically served rural areas where they did not have physical presence by running periodic eye camps. Skilled staff traveled to the eye camps and provided villagers with eye care about twice a year. The camps could serve only a small fraction of the needs of the rural areas (less than 10% according to AEH). Given the availability of a communications network with high quality connectivity, AEH altered its basic service delivery model for rural areas. Remote vision centers were set up closer to the rural population. These were staffed by a paramedic staff trained by AEH to perform basic standardized tasks that were traditionally handled by an opthalmologist. Using a video conferencing link, complex diagnosis was handled by an opthalmologist stationed at one of the main hospitals. A single opthalmologist could now serve multiple vision centers ensuring a high level of utilization. By moving simpler, standardized tasks closer to rural patients while centralizing the expensive and highly skilled opthalmologists, AEH lowered the additional expense incurred while significantly increasing coverage and reducing the travel time for the patient. AEH reported that the telemedicine-based model allowed it to "increase access to high quality eye care by over ten times while reducing the cost of access for poor patients ten fold."

Summary

Chapter 1 stated that the effectiveness of any process depends on its current and past performance and on the future goals as expressed by the firm's strategy. In this chapter, we focused on the relationship between strategy and process.

Strategic positioning means deliberately performing activities different from or better than those of the competition. Operations strategy consists of plans to develop the desired process competencies. Operational effectiveness requires developing processes and operating policies that support the strategic position better than the competitors. Both strategic positioning and operational effectiveness are necessary for gaining and sustaining competitive advantage.

The key insight is that an effective business process is tailored to its business strategy—process structure and operating policies work together to support the organization's overall strategic objectives. To ensure such strategic fit, a three-step approach can be adopted. First, determine the strategic positioning by prioritizing the targeted customer needs of product cost, quality, variety, and response time. Second, determine what the process should be good at to support that strategic position: In other words, infer the necessary process competencies in terms of process cost, quality, flexibility, and flow time. Finally, given that different processes have different competencies, design a process whose competencies best support the strategy.

Focusing operations and matching products with processes are means of facilitating an effective fit between strategy and processes. Given that the best operational practices improve constantly in competitive industries, firms must make continuous improvements in their processes to maintain operational effectiveness.

Key Terms

- American system of manufacturing
- Business strategy
- Corporate strategy
- Competitive product space
- Cost efficiency
- Division of labor
- Enterprise resource planning
- Flexible mass production
- Focused process
- Focused strategy
- Functional specialization
- Functional strategies
- Information Technology
- Market-driven strategy
- Mass production
- Operational effectiveness
- Operations frontier
- Operations strategy
- Plant-within-a-plant (PWP)
- Process-driven strategy
- Productivity dilemma
- Product–process matrix
- Product specialization
- Statistical quality control
- Strategic fit
- Strategic positioning
- Trade-off

Discussion Questions

2.1 How do the strategies of your neighborhood supermarket differ from those of Walmart? How do their business processes support those strategies?

2.2 Compare and contrast the strategies and supporting business processes of Southwest Airlines and Singapore Airlines. That is, do some research (e.g., read their corporate Web sites) and compare their business strategy in terms of the four product dimensions, targeted market segments, and their process architectures.

2.3 Consider the life cycle of a product from introduction to maturity to decline. What kind of process (job shop, batch process, or line flow) would be appropriate at each stage, and why?

2.4 A small printed-circuit-board manufacturer has established itself on the ability to supply a large variety of small orders in a timely manner. A firm approaches the manufacturer to place a large order (equal to the current volume from all other orders) but asks for a 30 percent discount. Do you think the circuit board manufacturer should accept the order? Justify your answer.

2.5 Compare the differences in patient needs at an emergency room in a hospital with that of a department doing knee replacements in terms of price, quality, time, and variety. Which of these departments should follow the approach taken by Shouldice? Why?

2.6 Briefly give an argument supporting the claim that "the essence of strategy is choosing what to do and what *not* to do."

2.7 MDVIP is a group of primary care physicians in Florida that offers a unique pricing structure. They charge patients a fixed fee of $1,500 per year on top of fees per visit over the year. That is, patients pay $1,500 even if they do not see a doctor during the year and still pay per consultation despite having paid the annual fee. The per-visit fees are comparable to industry averages and are covered by patients' health insurance. The fixed fee is not covered by standard health insurance. Since introducing the pricing format, the size of MDVIP's practice has shrunk from 6,000 patients to 300. Doctor's schedules are consequently more open, making it easier for patients to schedule appointments. In addition, doctor–patient consultations are well above the industry average of just under ten minutes.

Using the four competitive dimensions, how would you describe MDVIP's strategic position relative to a comparable, traditional practice? What are the implications for how MDVIP must manage its resources?

2.8 There are a surprising number of styles of baby strollers and carriages available in the marketplace. One style is the jogging stroller, which features oversize wheels that make the stroller easy to push on rough surfaces as the parent jogs. One maker of jogging strollers is Baby Trend. Baby Trend makes a full line of baby products—from diaper pails to high chairs to a variety of strollers. Their jogging strollers consist of a cloth seat stretched over a metal frame, a fairly standard design in the industry. They make a limited number of styles of single joggers meant to carry one baby (mostly differentiated by the color of the cloth) and one style of double stroller meant to carry two children. In the Baby Trend double stroller, the children sit next to each other (again a standard industry design). Baby Trend sells through independent Web sites (e.g., www.strollers4less.com), specialty stores (e.g., The Right Start), and "big box" retailers (e.g., Toys "R" Us). Their prices range from under $100 for a simple single jogger to $299 for their double stroller.

Another competitor is Berg Design. Jogging strollers are the only baby products Berg Design makes, although it produces other products that involve metalworking. Berg Design's joggers offer a

unique design in which the child sits in a molded plastic seat bolted to a metal frame. With a single seat, the child can face forward or backward (i.e., looking at mom or dad). For their multiple-seat strollers, the children sit in tandem as opposed to side by side. On the two-seat model, the children can both face forward or can face each other. They make models that can handle up to four children that are popular with day care centers. Berg Design's Web site emphasizes that each jogger "is made one at a time, no shortcuts; hand welded, powder painted and assembled with the utmost attention to craftsmanship." Berg Design sells directly to customers through the Web and a toll free number. Their prices range from $255 for their cheapest single jogger to $745 for a four-seat model.

a. How would you describe the strategic positions of Baby Trend and Berg Design?

b. How would you expect Baby Trend and Berg Design to have structured their respective processes for building strollers?

2.9 Give two main benefits of focus. Provide at least two reasons all organizations do not employ focused operations. (Note: Claiming a firm is ignorant of focus is not an option.)

Selected Bibliography

Ambrose, S. E. *Undaunted Courage: Meriwether Lewis Thomas Jefferson and the Opening of the American West*. New York: Simon & Schuster, 1996.

Anupindi, R., M. Aundhe, and M. Sarkar. "Healthcare Delivery Models and the Role of Telemedicine." In *Indian Economic Superpower: Fiction or Future*. Singapore: World Scientific, 2009.

Brandes, D. *Konsequent Einfach: Die Aldi-Erfolgsstory*. Frankfurt: Campus Verlag, 1998.

Billington, C., and F. Jager. "Procurement: The Missing Link in Innovation." *Supply Chain Management Review*, no. 1 (January–February 2008): 22–28.

Carpenter, M. "Dell Computers Founder Tells His Success Story." *The Diamondback Online*, April 7, 2003. www.inform.umd.edu/News/Diamondback/archives/2003/04/07/news6.html (last accessed September 2004).

Chopra, S., and J. A. Van Mieghem.. "Which E-Business Is Right for Your Supply Chain?" *Supply Chain Management Review* 4, no. 3 (July–August 2000): 32–40.

Collis, D. J., and C. A. Montgomery. "Competing on Resources: Strategy in the 1990s." *Harvard Business Review* 73, no. 4 (July–August 1995): 118–129.

Harrison, J. M., and C. Loch. "Five Principles of Business Process Reengineering." Unpublished manuscript. 1995.

Hayes, R. H., and S. C. Wheelwright. "Link Manufacturing Process and Product Life Cycles." *Harvard Business Review* 57, no. 1 (January–February 1979): 133–140.

Heskett, J. L. *Shouldice Hospital Limited*. Harvard Business School Case Study 9-683-068. Cambridge, Mass.: Harvard Business School, 1989. 1–16.

Hindle, T. *Field Guide to Strategy*. Cambridge, Mass.: Harvard Business School Press, 1994.

Hounshell, D. A. *From the American System to Mass Production 1800–1932: The Development of Manufacturing Technology in the United States*. Baltimore: Johns Hopkins University Press, 1984.

Jaikumar, R. "Postindustrial Manufacturing." *Harvard Business Review* 64, no. 6 (November–December 1986): 69–75.

Kanigel, R. "The One Best Way: Frederick Winslow Taylor and the Enigma of Efficiency." Viking: New York, 1997.

Maguire, J. "Insights-Trends: Case Study: Walmart.com." November 15, 2002. www.ecommerce-guide.com/news/trends/article.php/10417_1501651 (accessed April 19, 2011).

Porter, M. E. "What Is Strategy?" *Harvard Business Review* 74, no. 6 (November–December 1996): 61–78.

Rubin, H. "The Perfect Vision Dr. V." *Fast Company*, no. 43 (February 2001): 146.

Skinner, W. "Manufacturing—Missing Link in Corporate Strategy." *Harvard Business Review* 47, no. 3 (May–June 1969): 136–145.

Skinner, W. "The Focused Factory." *Harvard Business Review* 52, no. 3 (May–June 1974): 113–121.

Stalk, G., P. Evans, and L. E. Shulman. "Competing on Capabilities: The New Rules of Corporate Strategy." *Harvard Business Review* 70, no. 2 (March–April 1992): 54–69.

Stalk, G., and A. M. Webber. "Japan's Dark Side of Time." *Harvard Business Review* 71, no. 4 (July–August 1993): 93–101.

Treacy, M., and F. Wiersema. *The Discipline of Market Leaders*. Boston: Addison-Wesley, 1997.

Van Mieghem, J. *Operations Strategy: Principles and Practice*. Belmont, MA: Dynamic Ideas, 2008.

Visnic, B. "Toyota Adopts New Flexible Assembly System." *Ward's Autoworld* (November, 2002): 30–31.

This page intentionally left blank

PART II

Process Flow Metrics

Chapter 3

Process Flow Measures

INTRODUCTION

Vancouver International Airport Authority manages and operates the Vancouver International Airport. Its focus on safety, security, and customer service has contributed to Vancouver International Airport being named the winner of best airport in North America award at the 2010 Skytrax World Airport Awards. In order to maintain its excellent customer service standards and in anticipation of new government regulations, airport management wanted to reduce the time customers spent in the airport security checkpoints. They wanted to improve the way that customers flowed through the process. In other words, they sought to better their *process flow.*

BellSouth International is a provider of wireless services in 11 Latin American countries. As a service provider, the company leases its network capacity on a monthly basis to two categories of customers: prepaid and postpaid. One of the most time-consuming processes for the company in the Latin American market is the service activation process: getting a wireless telephone into the hands of interested potential customers.

The various steps in the activation process include determination of the type of wireless service, credit check, selection of phone and service plan, assignment of the phone number, making a test call, and providing a short tutorial. At one of its largest activation

centers, the company serves an average of 10,000 customers per week with 21 percent being activated with a postpaid account and the remaining with a prepaid account.

To manage and improve this activation process, the following questions must be answered: What operational measures should a manager track as leading indicators of the financial performance of the process? How does the time to process a customer and the number of customers that are being served at any point in time impact the total number of customers that can be served per week? How do these process measures impact the financial performance of the process? Which specific outcomes can be called "improvements?" How can we prioritize our improvements into an executable action plan?

This chapter aims to provide answers to these questions. We will define process flow in Section 3.1 of this chapter. Then, in Sections 3.2 through 3.4, we will introduce the three fundamental measures of process performance: inventory, throughput, and flow time. In Section 3.5, we will explore the basic relationship among these three measures, called Little's law. Section 3.6 shows how Little's law can be used to analyze financial statements. We will discuss the related concepts of takt time and inventory turns in Section 3.7. Finally, Section 3.8 links these process flow measures to financial measures of performance to determine when a process change (e.g., reengineered process flows or allocation of additional resources) has been an improvement from both operational and financial perspectives.

3.1 THE ESSENCE OF PROCESS FLOW

Thus far, we have learned that the objective of any process is to transform inputs into outputs (products) to satisfy customer needs. We also know that while an organization's strategic position establishes *what* product attributes it aims to provide, its operational effectiveness measures *how well* its processes perform this mission. We have seen that product attributes and process competencies are classified in terms of cost, time, variety, and quality. We noted that, to improve any process, we need internal measures of process performance that managers can control. We also saw that if chosen carefully, these internal measures can serve as leading indicators of customer satisfaction and financial performance as well.

In this chapter we focus on **process flow measures**—*three key internal process performance measures that together capture the essence of process flow: flow time, flow rate, and inventory*. As we will see in subsequent chapters, these three process-flow measures directly affect process cost and response time, and they are affected by process flexibility (or lack thereof) and process quality.

Throughout this book, we examine processes from the perspective of *flow.* Specifically, we look at the process dynamics as inputs enter the process, flow through various activities performed (including such "passive activities" as waiting for activities to be performed), and finally exit the process as outputs. Recall from Chapter 1 that a flow unit is a unit flowing through a process. A flow unit may be a patient, a dollar, a pound of steel, a customer service request, a research-and-development project, or a bank transaction to be processed. In our study of process flow performance, we look at three measures and answer three important questions:

1. On average, how much time does a typical flow unit spend within the process boundaries?
2. On average, how many flow units pass through the process per unit of time?
3. On average, how many flow units are within the process boundaries at any point in time?

The case of Vancouver International Airport, described in Example 3.1, is an example of a business situation in which examining process flow performance is particularly

useful. Later in this chapter, we will analyze this example to determine whether a process change leads to an improvement.

EXAMPLE 3.1

Now, let us begin to look at how the Vancouver International Airport Authority went about improving its customer flow through its airport security checkpoints. To understand customer flow, managers began by analyzing a single security screening line, which is comprised of an X-ray scanner with an operator and screening officers. Arriving customers either queue up or, if there is no queue on arrival, directly put their bags on the scanner. While customers can have 0, 1, 2, or 3 carry-on bags, including purses, wallets, and so on, on average, a typical customer has 1.5 bags. The X-ray scanner can handle 18 bags per minute. On average, about half the passengers arrive at the checkpoint about 40 minutes (±10 minutes) before departure for domestic flights. The first passenger shows up about 80 minutes before departure, and the last passenger arrives 20 minutes before departure. For a flight with 200 passengers, this gives the following approximate arrival rate pattern: About 75 passengers arrive 80 to 50 minutes early, 100 arrive 50 to 30 minutes early, and the remaining 25 arrive between 30 to 20 minutes before scheduled departure.

To minimize layover time for passengers switching flights, many of Vancouver's flights depart around the same time. As we look at the three key process measures, we will assume for simplicity that exactly three flights, each carrying 200 passengers, are scheduled for departure each hour: that is, three flights depart at 10 A.M., three flights at 11 A.M., and so forth. With increased security procedures, however, the simultaneous departures of flights were overwhelming scanner capacity and creating long waiting times. The airport authority needed to know how staggering flight departures—for example, spreading out departures so that one flight would depart every 20 minutes—would affect the flow and waiting times of passengers through the security checkpoint.

3.2 THREE KEY PROCESS MEASURES

Flow Time Recall from Chapter 1 that processes transform flow units through networks of activities and buffers. Thus, as a flow unit moves through the process, one of two things happens to it:

1. It undergoes an activity.
2. It waits in a buffer to undergo an activity.

In the airport example, passengers and their luggage are either security scanned or wait in line before the X-ray machine. Let us follow a specific passenger or flow unit from the time it enters the process until the time it exits. *The total time spent by a flow unit within process boundaries is called* **flow time**. Some flow units move through the process without any wait; perhaps they require only resources that are available in abundance (there are several X-ray scanners and operators available), or they arrive at times when no other flow units are present (there are no other passengers checking through security when they arrive), or they are artificially expedited (a first-class passenger conceivably could be given priority over economy-class passengers). Others, meanwhile, may spend a long time in the process, typically waiting for resources to become available. In general, therefore, flow time varies—sometimes considerably—from one flow unit to another.

As a measure of process performance, flow time indicates the time needed to convert inputs into outputs and includes any time spent by a flow unit waiting for processing

activities to be performed. It is thus useful information for a manager who must promise a delivery date to a customer. It also indicates how long working capital, in the form of inventory, is tied up in the process.

Flow Rate An important measure of process flow dynamics is *the number of flow units that flow through a specific point in the process per unit of time*, which is called **flow rate**. In many settings, flow rates may change over time so that in addition to the specific point in the process, we also must specify the time when the measurement was taken. In Example 3.1, the inflow rate of passengers at a security checkpoint at Vancouver International Airport changes over time. Recall that for each of the three flights, about half the 200 passengers for each flight arrive between 50 and 30 minutes before flight departure. So for each of the three flights departing at 10 A.M., about 100 passengers arrive between 9:10 and 9:30 A.M., a 20-minute interval. This means that a total of 300 passengers arrive during this time period for the three flights, giving an inflow rate of roughly 15 passengers per minute. The remaining 300 passengers for the three flights arrive between 8:40 and 9:10 A.M. (about 80 to 50 minutes before departure) and between 9:30 and 9:40 A.M. (about 30 to 20 minutes before departure). That is, the remaining 300 passengers arrive during a total time period of 40 minutes, giving an inflow rate of 7.5 passengers per minute, which is half the inflow rate during the peak time period from 9:10 to 9:30 A.M. The outflow rate of the checkpoint, however, is limited by the X-ray scanner, which cannot process more than 18 bags per minute or, with an average of 1.5 bags per passenger, 12 passengers per minute.

When we consider the flow rate at a specific point in time t, we call it the *instantaneous* flow rate and denote it by $R(t)$. For example, if we focus on the flow through entry and exit points of the process at time t, we can denote the instantaneous total inflow and outflow rates through all entry and exit points, respectively, as $R_i(t)$ and $R_o(t)$.

The process that is shown graphically in Figure 3.1 features two entry points and one exit point. Total inflow rate $R_i(t)$, then, is the sum of the two inflow rates, one each from the two entry points. Remember that inputs may enter a process from multiple points and that outputs may leave it from multiple points.

Inventory When the inflow rate exceeds the outflow rate, the number of flow units inside the process increases. Inventory is the total number of flow units present within process boundaries. In the airport example, during the peak period of 9:10 to 9:30 A.M., the inflow rate is 15 passengers per minute, while the outflow rate is 12 passengers per minute. Hence, an inventory of passengers will build in the form of a queue. We define the total number of flow units present within process boundaries at time t as the *process inventory at time t* and denote it by $I(t)$. To measure the process inventory at time t, we take a snapshot of the process at that time and count all the flow units within process boundaries at that moment. Current inventory thus represents all flow units that have entered the process but have not yet exited.

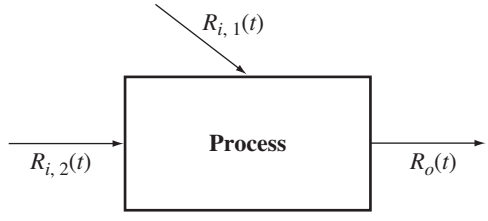


FIGURE 3.1 Input and Output Flow Rates for a Process with Two Entry Points

Inventory has traditionally been defined in a manufacturing context as material waiting to be processed or products waiting to be sold. Our definition considers a general flow unit and thus takes a much broader view that applies to any process, whether it is a manufacturing, a service, a financial, or even an information process. Inventory can thus encompass products, customers, cash, and orders. Our definition of inventory includes all flow units within process boundaries—whether they are being processed or waiting to be processed. Thus, raw materials, work in process (partially completed products), and finished goods inventories are included. This broader definition of inventory allows us to provide a unified framework for analyzing flows in all business processes.

What constitutes a flow unit depends on the problem under consideration. By defining the flow unit as money—such as a dollar, a euro, or a rupee—we can analyze financial flows. Adopting money as the flow unit and our broader view of inventory, we can use inventory to identify the working capital requirements. A key financial measure for any process is investment in working capital. Accountants define working capital as current assets minus current liabilities. Current assets include the number of dollars within process boundaries in the form of inventory as well as in the form of cash and any accounts receivable. Thus, inventory is like money that is tied up: A reduction in inventory reduces working capital requirements. Reduced working capital requirements reduce the firm's interest expense or can make extra cash available for investment in other profitable ventures. (Reducing inventory also reduces flow time and improves responsiveness, as we shall see later in this chapter.)

3.3 FLOW TIME, FLOW RATE, AND INVENTORY DYNAMICS

Generally, both inflow and outflow rates fluctuate over time. When the inflow rate exceeds the outflow rate in the short term, the inventory increases, or builds up. In contrast, if outflow rate exceeds inflow rate in the short term, the inventory decreases. Thus, inventory dynamics are driven by the difference between inflow and outflow rates. We define the **instantaneous inventory accumulation (buildup) rate**, $\Delta R(t)$, as *the difference between instantaneous inflow rate and outflow rate:*

$$\text{Instantaneous inventory accumulation (or buildup) rate } \Delta R(t) = \text{Instantaneous inflow rate } R_i(t) - \text{Instantaneous outflow rate } R_o(t)$$

or

$$\Delta R(t) = R_i(t) - R_o(t) \qquad \textbf{(Equation 3.1)}$$

Thus, the following holds:

- If instantaneous inflow rate $R_i(t) >$ instantaneous outflow rate $R_o(t)$, then inventory is accumulated at a rate $\Delta R(t) > 0$.
- If instantaneous inflow rate $R_i(t) =$ instantaneous outflow rate $R_o(t)$, then inventory remains unchanged.
- If instantaneous inflow rate $R_i(t) <$ instantaneous outflow rate $R_o(t)$, then inventory is depleted at a rate $\Delta R(t) < 0$.

For example, if we pick a time interval (t_1, t_2) during which the inventory buildup rate ΔR is constant, the associated change in inventory during that period is

$$\text{Inventory change} = \text{Buildup rate} \times \text{Length of time interval}$$

or

$$I(t_2) - I(t_1) = \Delta R \times (t_2 - t_1) \qquad \textbf{(Equation 3.2)}$$

Given an initial inventory position and dividing time into intervals with constant accumulation rates, we can construct an **inventory buildup diagram** that depicts *inventory fluctuation over time.* On the horizontal axis we plot time, and on the vertical axis we plot the inventory of flow units at each point in time. Assuming that we start with zero inventory, the inventory at time t is the difference between the cumulative inflow and outflow up to time t. Example 3.2 provides an illustration of an inventory buildup diagram.

EXAMPLE 3.2

MBPF Inc. manufactures prefabricated garages. The manufacturing facility purchases sheet metal that is formed and assembled into finished products—garages. Each garage needs a roof and a base, and both components are punched out of separate metal sheets prior to assembly. Production and demand data for the past eight weeks are shown in Table 3.1. Observe that both production and demand vary from week to week.

We regard the finished goods inventory warehouse of MBPF Inc. as a process and each garage as a flow unit. The production rate is then the inflow rate, while demand (sales) is the outflow rate. Clearly, both have fluctuated from week to week.

MBPF Inc. tracks inventory at the end of each week, measured in number of finished garages. Let $I(t)$ denote the inventory at the end of week t. Now suppose that starting inventory at the beginning of week 1 (or the end of week 0) is 2,200 units, so that

$$I(0) = 2{,}200$$

Now, subtracting week 1's production or inflow rate $R_i(1) = 800$ from its demand or outflow rate $R_o(1) = 1{,}200$ yields an inventory buildup rate:

$$\Delta R(1) = 800 - 1{,}200 = -400 \text{ for week 1}$$

So, the ending inventory at week 1 is

$$I(1) = I(0) + \Delta R(1) = 2{,}200 + (-400) = 1{,}800$$

We can similarly evaluate buildup rates and inventory for each week, as shown in Table 3.1. Clearly, the inventory of flow units varies over time around its average of 2,000 garages.

With these data, we can construct an inventory buildup diagram that depicts how inventory fluctuates over time. Figure 3.2 shows the inventory buildup diagram for MBPF over the eight weeks considered, where we have assumed, for simplicity, that inventory remains constant during the week and changes only at the end of the week when sales take place.

Table 3.1 Production, Demand, Buildup Rate, and Ending Inventory for MBPF Inc.

Week	0	1	2	3	4	5	6	7	8	Average
Production		800	1,100	1,000	900	1,200	1,100	950	950	1,000
Demand		1,200	800	900	1,100	1,300	1,300	550	850	1,000
Buildup rate ΔR		−400	300	100	−200	−100	−200	400	100	0
Ending inventory	2,200	1,800	2,100	2,200	2,000	1,900	1,700	2,100	2,200	2,000

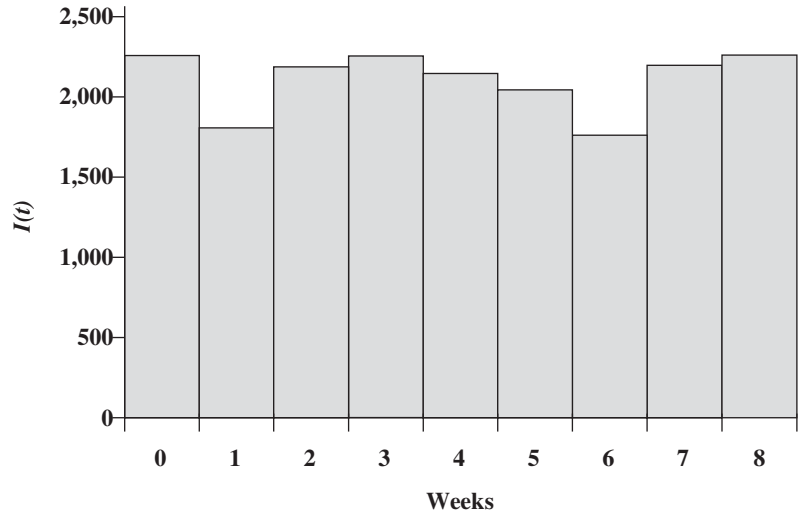


FIGURE 3.2 Inventory Buildup Diagram for MBPF Inc.

We can also analyze the airport security process of Example 3.1 by deriving the build-up of an inventory of passengers (and their associated waiting times) from the flow accumulation rate. We will define the X-ray scanner as our process and consider the representative example of three flights that are scheduled to depart at 10 A.M. Assume that we start at 8:40 A.M., when no passengers are waiting in line.

As derived earlier, the inflow rate during 8:40 to 9:10 A.M. is 7.5 passengers per minute. The outflow rate from the queue is the rate at which baggage is scanned. While the X-ray scanner can process up to 12 passengers per minute, it cannot process more passengers than are arriving, so the outflow rate also is 7.5 per minute. Thus, as summarized in Table 3.2, from 8:40 to 9:10 A.M., the buildup rate in the queue is zero: The X-ray scanner can easily keep up with the inflow, and no passengers have to wait in line. During the peak arrival period of 9:10 to 9:30 A.M., however, the inflow rate of 15 passengers per minute exceeds the maximal scanner outflow rate of 12 passengers per minute, so that a queue (inventory) starts building at $\Delta R = 3$ passengers per minute. At 9:30 A.M., the line for the scanner has grown to $\Delta R \times (9{:}30 - 9{:}10) = 3$ passengers per minute $\times$ 20 minutes = 60 passengers! After 9:30 A.M., the X-ray scanner keeps processing at the full rate of 12 passengers per minute, while the inflow rate has slowed to the earlier lower rate of 7.5 passengers per minute, so that the passenger queue is being depleted at a rate of 4.5 passengers per minute. Thus, the 60-passenger queue is eliminated after $60/4.5 = 13.33$ minutes. In other words, at 9:43 and 20 seconds, the queue is empty again, and the X-ray scanner can keep up with the inflow.

Table 3.2 Buildup Rates and Ending Inventory Data: Vancouver Airport Security Checkpoint of Example 3.1

Time	8:40 A.M.	8:40–9:10 A.M.	9:10–9:30 A.M.	9:30–9:43:20 A.M.	9:43:20–10:10 A.M.
Inflow rate R_i		7.5/min.	15/min.	7.5/min.	7.5/min.
Outflow rate R_o		7.5/min.	12/min.	12/min.	7.5/min.
Buildup rate ΔR		0	3/min.	4.5/min.	0
Ending inventory (number of passengers in line)	0	0	60	0	0

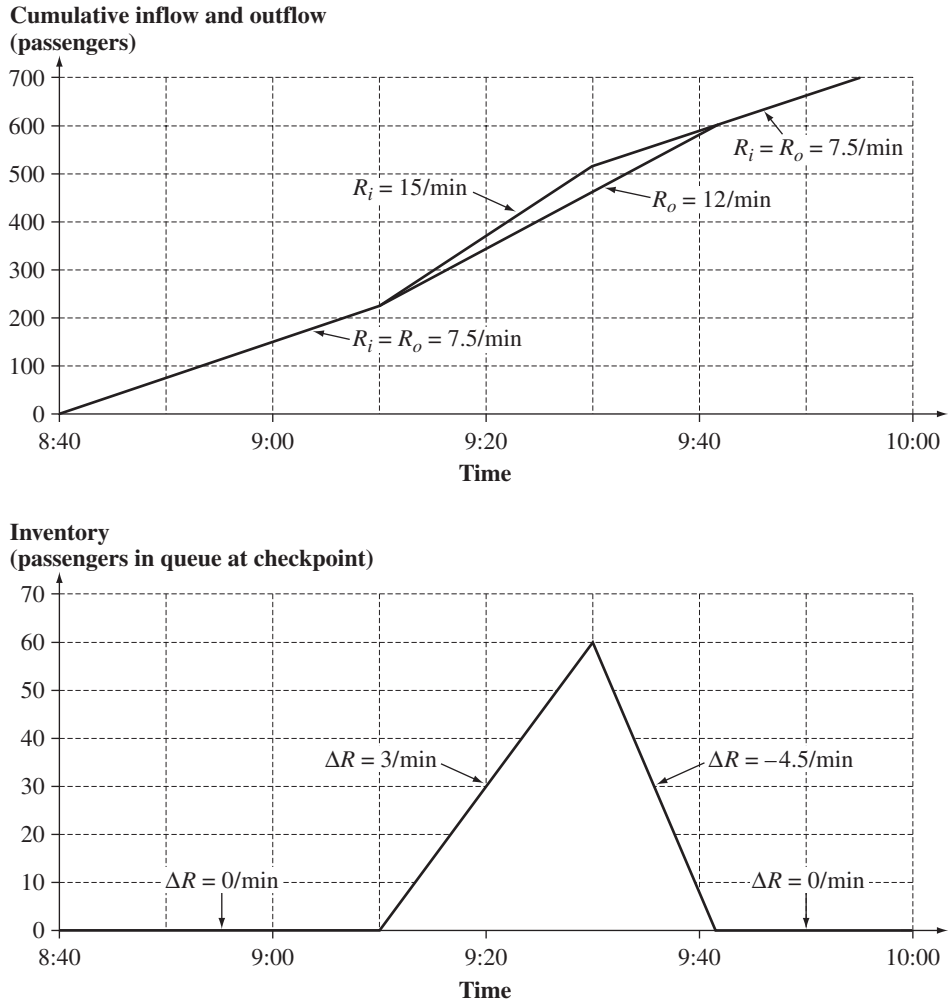


FIGURE 3.3 Inventory Buildup Diagram for Vancouver Airport Security Checkpoint

Observe that while the lower inflow rate of 7.5 passengers per minute for the 10 A.M. flights ends at 9:40 A.M., the first set of passengers start arriving for the 11 A.M. flights at 9:40 A.M. Just as the queue starts building up at 9:10 A.M. for the 10 A.M. flight, it will start building up again at 10:10 A.M. for the 11 A.M. flights. Thus, the cycle repeats itself for the next set of flight departures. Figure 3.3 shows the inventory buildup diagram together with the associated cumulative number of passengers arriving to and departing from the checkpoint process and clearly indicates that the difference between cumulative inflows and outflows is inventory.

Now, if flight departures are staggered, the peaks and valleys in arrival rates for different flights cancel each other out, as illustrated in Table 3.3. (The shaded time buckets correspond to the passenger arrivals for a particular flight.) Spreading out flight departures thus gives a constant arrival rate of 600 passengers per hour, which equals 10 passengers per minute at any point in time throughout the day. This is well below the process capacity of the X-ray scanner, so that the buildup rate would be zero. In short, by staggering the flights, no queues would develop at the security checkpoint. (The previous analysis is approximate because it ignores the short-time variability of passenger arrivals within any small time interval. This impact of short-time variability on process performance will be discussed in Chapter 8.)

Table 3.3 Inflow Rates with Staggered Departures for Vancouver Airport Security Checkpoint of Example 3.1

Time (start of period)	**8:40**	**8:50**	**9:00**	**9:10**	**9:20**	**9:30**	**9:40**	**9:50**	**10:00**	**10:10**	**10:20**	**10:30**
Inflow rate for flights arriving on the hour, $R_{i,1}$	2.5/min.	2.5/min.	2.5/min.	5/min.	5/min.	2.5/min.	2.5/min.	2.5/min.	2.5/min.	5/min.	5/min.	2.5/min.
Inflow rate for flights arriving 20 minutes past the hour, $R_{i,2}$	5/min.	2.5/min.	2.5/min.	2.5/min.	2.5/min.	5/min.	5/min.	2.5/min.	2.5/min.	2.5/min.	2.5/min.	5/min.
Inflow rate for flights arriving 40 minutes past the hour, $R_{i,3}$	2.5/min.	5/min.	5/min.	2.5/min.	2.5/min.	2.5/min.	2.5/min.	5/min.	5/min.	2.5/min.	2.5/min.	2.5/min.
Total inflow rate R_i	10/min.	10/min.	10/min.	10/min.	10/min.	10/min.	10/min.	10/min.	10/min.	10/min.	10/min.	10/min.

3.4 THROUGHPUT IN A STABLE PROCESS

A **stable process** *is one in which, in the long run, the average inflow rate is the same as the average outflow rate.* In the airline checkpoint example, while inflow and outflow rates change over time, the *average* inflow rate is 600 passengers per hour. Because the X-ray scanner can process up to 12 passengers per minute, equaling 720 passengers per hour, it can easily handle the inflow over the long run so that the *average* outflow rate also is 600 passengers per hour. Thus, the security checkpoint with the unstaggered flights is a stable process.

When we have a stable process, we refer to average inflow or outflow rate as **average flow rate**, or **throughput**, which is *the average number of flow units that flow through (into and out of) the process per unit of time.* We will denote the throughput simply as R to remind ourselves that throughput is a rate. As a measure of process performance, throughput rate tells us the average rate at which the process produces and delivers output. Ideally, we would like process throughput rate to match the rate of customer demand. (If throughput is less than the demand rate, some customers are not served. If the converse is true, the process produces more than what is sold.)

Consider the inventory dynamics *of the original situation* in a stable process we can define the average inventory over time and denote this by I. For example, let us find the average inventory at the Vancouver International Airport. Consider the inventory dynamics as shown in the bottom picture of Figure 3.3. From 8:40 to 9:10 A.M., the inventory or queue before the airline checkpoint is zero. From 9:10 through 9:43 A.M., the inventory builds up linearly to a maximum size of 60 and then depletes linearly to zero. Thus, the average inventory during that period is $60/2 = 30$. (Recall that the average height of a triangle is half its maximum height.) Finally, the inventory is zero again from 9:43 to 10:10 A.M., when the cycle repeats with the next inventory buildup. To estimate the average queue size, it is then sufficient to consider the 60-minute interval between the start of two consecutive inventory buildups; for example, from 9:10 A.M. (when inventory builds up for the 10 A.M. flights) to 10:10 A.M. (when inventory starts to build up for 11 A.M. flights). As we have seen, during this interval there is an average of 30 passengers between 9:10 and 9:43 A.M. and zero passengers between 9:43 and 10:10 A.M. Thus, the average queue size is the time-weighted average:

$$I = \frac{33 \text{ min.} \times 30 \text{ passengers} + 27 \text{ min.} \times 0 \text{ passengers}}{60 \text{ min.}}$$
$$= 16.5 \text{ passengers}$$

While the average inventory accumulation rate ΔR must be zero in a stable process (remember, average inflow rate equals average outflow rate), the average inventory, typically, is positive.

Now, let us look at **average flow time**. While the actual flow time varies across flow units, we can define the average flow time as *the average (of the flow times) across all flow units that exit the process during a specific span of time.* We denote the average flow time by T. One method to measure the average flow time is to track the flow time of each flow unit over a long time period and then compute its average. Another method is to compute it from the throughput and the average inventory, which we will explain next.

3.5 LITTLE'S LAW: RELATING AVERAGE FLOW TIME, THROUGHPUT, AND AVERAGE INVENTORY

The three performance measures that we have discussed answer the three questions about process flows that we raised earlier:

1. On average, how much time does a typical flow unit spend within process boundaries? The answer is the *average flow time T*.

2. On average, how many flow units pass through the process per unit of time? The answer is the *throughput R*.
3. On average, how many flow units are within process boundaries at any point in time? The answer is the *average inventory I*.

Little's Law We can now show that in a stable process, there is a fundamental relationship among these three performance measures. This relationship is known as **Little's law**, which states that *average inventory equals throughput multiplied by average flow time:*

$$\text{Average Inventory } (I) = \text{Throughput } (R) \times \text{Average Flow Time } (T)$$

or

$$I = R \times T \qquad \textbf{(Equation 3.3)}$$

To see why Little's law must hold, let us mark and track an arbitrary flow unit. After the marked flow unit enters the process boundaries, it spends T time units before departing. During this time, new flow units enter the process at rate R. Thus, during the time T that our marked flow unit spends in the system, $R \times T$ new flow units arrive. Thus, at the time our marked flow unit exits the system, the inventory is $R \times T$. Because our marked flow unit was chosen at random and because the process is stable, the average inventory within process boundaries that a randomly picked flow unit sees, I, must be the same as $R \times T$.

Little's law allows us to derive the flow time averages of all flow units from the average throughput and inventory (which are averages over time and typically easier to calculate than average flow units). In the airport security checkpoint example, we found that average queue size $I = 16.5$ passengers, while throughput was $R = 600$ passengers per hour $= 10$ passengers per minute. To determine the average time spent by a passenger in the checkpoint queue, we use Little's law, $I = R \times T$ and solve for T so that

$$T = I/R = 16.5 \text{ passengers}/10 \text{ passengers per minute} = 1.65 \text{ minutes}$$

Recall that many passengers do not wait at all, while the passenger who waits longest is the one who arrives when the queue is longest at 60 passengers. That unfortunate passenger must wait for all 60 passengers to be processed, which implies a waiting time of 60/12 minutes = 5 minutes. Example 3.3 illustrates Little's law for the MBPF Inc. example.

EXAMPLE 3.3

Recall that average inventory at MBPF Inc. in Example 3.2 was $I = 2000$ garages. Computing the average production over the eight weeks charted in Table 3.1 yields an average production rate of 1,000 garages per week. Average demand experienced by MBPF Inc. over the eight weeks considered in Table 3.1 is also 1,000 garages. Over the eight weeks considered, therefore, average production at MBPF has matched average demand. Because these rates are equal, we conclude that MBPF Inc. is a stable process with a throughput of 1,000 garages per week.

Now suppose that in terms of material and labor, each garage costs $3,300 to produce. If we consider each dollar spent as our flow unit, MBPF Inc. has a throughput of $R = \$3{,}300 \times 1{,}000$ garages $= \$3{,}300{,}000$ per week. Thus, we have evaluated the throughput rate of MBPF Inc. using two different flow units: garages and dollars. Similarly, I, inventory can be evaluated as 2,000 garages, or $2{,}000 \times \$3{,}300$ (the cost of each garage) $= \$6{,}600{,}000$.

Because this is a stable process, we can apply Little's law to yield the average flow time of a garage, or of a dollar tied up in each garage, as

$$T = I/R = \$6{,}600{,}000/\$3{,}300{,}000 = 2 \text{ weeks}$$

Two immediate but far-reaching implications of Little's law are the following:

1. Of the three operational measures of performance—average flow time, throughput, and average inventory—*a process manager need only focus on two measures because they directly determine the third measure from Little's law.* It is then up to the manager to decide which two measures should be managed.
2. For a given level of throughput in any process, the only way to reduce flow time is to reduce inventory and vice versa.

Now let us look at some brief examples that well illustrate the wide range of applications of Little's law in both manufacturing and service operations. It will be helpful to remember the following:

Average inventory is denoted by I.

Throughput is denoted by R.

Average flow time is denoted by T.

3.5.1 Material Flow

A fast-food restaurant processes an average of 5,000 kilograms (kg) of hamburgers per week. Typical inventory of raw meat in cold storage is 2,500 kg. The process in this case is the restaurant and the flow unit is a kilogram of meat. We know, therefore, that

$$\text{Throughput } R = 5{,}000 \text{ kg./week}$$

and

$$\text{Average inventory } I = 2{,}500 \text{ kg.}$$

Therefore, by Little's law,

$$\text{Average flow time } T = I/R = 2{,}500/5{,}000 = 0.5 \text{ week}$$

In other words, an average kilogram of meat spends only half a week in cold storage. The restaurant may use this information to verify that it is serving fresh meat in its hamburgers.

3.5.2 Customer Flow

The café Den Drippel in Ninove, Belgium, serves, on average, 60 customers per night. A typical night at Den Drippel is long, about 10 hours. At any point in time, there are, on average, 18 customers in the café. These customers are either enjoying their food and drinks, waiting to order, or waiting for their order to arrive. Since we would like to know how long a customer spends inside the restaurant, we are interested in the average flow time for each customer. In this example, the process is the café, the flow unit is a customer, and we know that

$$\text{Throughput } R = 60 \text{ customers/night}$$

$$\text{Since nights are 10 hours long, } R = 6 \text{ customers/hour}$$

and

$$\text{Average inventory } I = 18 \text{ customers}$$

Thus, Little's law yields the following information:

$$\text{Average flow time } T = I/R = 18/6 = 3 \text{ hours}$$

In other words, the average customer spends three hours at Den Drippel.

3.5.3 Job Flow

A branch office of an insurance company processes 10,000 claims per year. Average processing time is three weeks. We want to know how many claims are being processed at any given point. Assume that the office works 50 weeks per year. The process is a branch of the insurance company, and the flow unit is a claim. We know, therefore, that

$$\text{Throughput } R = 10{,}000 \text{ claims}/\text{year}$$

and

$$\text{Average flow time } T = 3/50 \text{ year}$$

Thus, Little's law implies that

$$\text{Average inventory } I = R \times T = 10{,}000 \times 3/50 = 600 \text{ claims}$$

On average, then, scattered in the branch are 600 claims in various phases of processing—waiting to be assigned, being processed, waiting to be sent out, waiting for additional data, and so forth.

3.5.4 Cash Flow

A steel company processes $400 million of iron ore per year. The cost of processing ore is $200 million per year. The average inventory is $100 million. We want to know how long a dollar spends in the process. The value of inventory includes both ore and processing cost. The process in this case is the steel company, and the flow unit is a cost dollar. A total of $400 million + $200 million = $600 million flows through the process each year. We know, therefore, that

$$\text{Throughput } R = \$600 \text{ million}/\text{year}$$

and

$$\text{Average inventory } I = \$100 \text{ million}$$

We can thus deduce the following information:

$$\text{Average flow time } T = I/R = 100/600 = 1/6 \text{ year} = 2 \text{ months}$$

On average, then, a dollar spends two months in the process. In other words, there is an average lag of two months between the time a dollar enters the process (in the form of either raw materials or processing cost) and the time it leaves (in the form of finished goods). Thus, each dollar is tied up in working capital at the factory for an average of two months.

3.5.5 Cash Flow (Accounts Receivable)

A major manufacturer bills $300 million worth of cellular equipment per year. The average amount in accounts receivable is $45 million. We want to determine how much time elapses from the time a customer is billed to the time payment is received. In this case, the process is the manufacturer's accounts-receivable department, and the flow unit is a dollar. We know, therefore, that

$$\text{Throughput } R = \$300 \text{ million}/\text{year}$$

and

$$\text{Average inventory } I = \$45 \text{ million}$$

Thus, Little's law implies that

$$\text{Average flow time } T = I/R = 45/300 \text{ year} = 0.15 \text{ year} = 1.8 \text{ months}$$

On average, therefore, 1.8 months elapse from the time a customer is billed to the time payment is received. Any reduction in this time will result in revenues reaching the manufacturer more quickly.

3.5.6 Service Flow (Financing Applications at Auto-Moto)

Auto-Moto Financial Services provides financing to qualified buyers of new cars and motorcycles. Having just revamped its application-processing operations, Auto-Moto Financial Services is now evaluating the effect of its changes on service performance. Auto-Moto receives about 1,000 loan applications per month and makes accept/reject decisions based on an extensive review of each application. We will assume a 30-day working month.

Until last year (under what we will call "Process I"), Auto-Moto Financial Services processed each application individually. On average, 20 percent of all applications received approval. An internal audit showed that, on average, Auto-Moto had about 500 applications in process at various stages of the approval/rejection procedure. In response to customer complaints about the time taken to process each application, Auto-Moto called in Kellogg Consultants (KC) to help streamline its decision-making process. KC quickly identified a key problem with the current process: Although most applications could be processed fairly quickly, some—because of insufficient and/or unclear documentation—took a disproportionate amount of time. KC, therefore, suggested the following changes to the process (thereby creating what we will call "Process II"):

1. Because the percentage of approved applications is fairly low, an Initial Review Team should be set up to preprocess all applications according to strict but fairly mechanical guidelines.
2. Each application would fall into one of three categories: *A* (looks excellent), *B* (needs more detailed evaluation), and *C* (reject summarily). *A* and *B* applications would be forwarded to different specialist subgroups.
3. Each subgroup would then evaluate the applications in its domain and make accept/reject decisions.

Process II was implemented on an experimental basis. The company found that, on average, 25 percent of all applications were *A*s, 25 percent *B*s, and 50 percent *C*s. Typically, about 70 percent of all *A*s and 10 percent of all *B*s were approved on review. (Recall that all *C*s were rejected.) Internal audit checks further revealed that, on average, 200 applications were with the Initial Review Team undergoing preprocessing. Just 25, however, were with the Subgroup A Team undergoing the next stage of processing and about 150 with the Subgroup B Team.

Auto-Moto Financial Services wants to determine whether the implemented changes have improved service performance.

Observe that the flow unit is a loan application. On average, Auto-Moto Financial Services receives and processes 1,000 loan applications per month. Let us determine the impact of the implemented changes on customer service.

Under Process I, we know the following:

$$\text{Throughput } R = 1{,}000 \text{ applications/month}$$

and

$$\text{Average inventory } I = 500 \text{ applications}$$

Thus, we can conclude that

$$\text{Average flow time } T = I/R$$

$$T = 500/1{,}000 \text{ month} = 0.5 \text{ month} = 15 \text{ days}$$

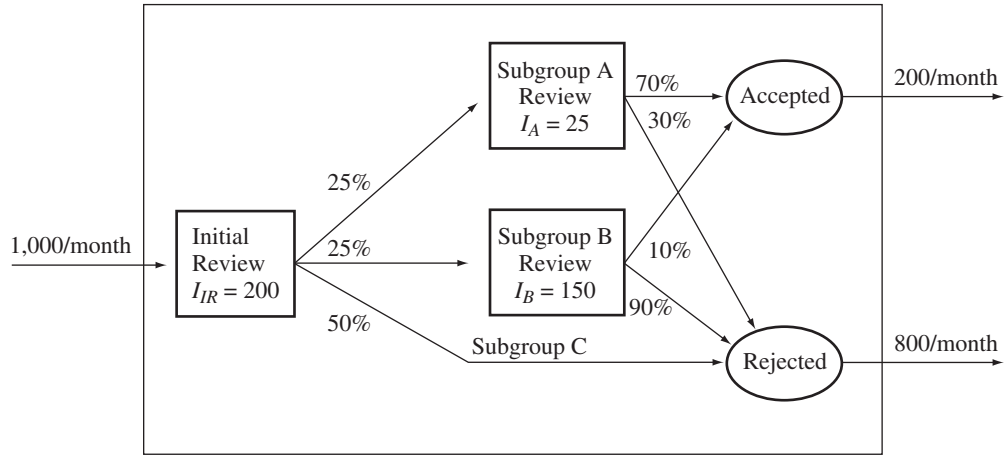


FIGURE 3.4 Flowchart for Auto-Moto Financial Services

In Process I, therefore, each application spent, on average, 15 days with Auto-Moto before receiving an accept/reject decision.

Now, let us consider Process II. Because this process involves multiple steps, it is better to start with the process flowchart in Figure 3.4. (We will discuss process flowcharts more fully in Chapter 4.) Note that, on average, 1,000 applications arrive per month for initial review. After initial review, 50 percent of these are rejected, 25 percent are categorized as Type A (looks excellent) and 25 percent are categorized as Type B (needs more detailed evaluation). On detailed evaluation by Subgroup A Team, 70 percent of Type A applications are accepted and 30 percent rejected. On evaluation by Subgroup B Team, 10 percent of Type B applications are accepted and 90 percent rejected. Thus, each month, an average of 200 applications is accepted and 800 rejected.

Furthermore, on average, 200 applications are with the Initial Review Team, 25 with the Subgroup A Team, and 150 with the Subgroup B Team. Thus, we can conclude that for Process II

$$\text{Throughput } R = 1{,}000 \text{ applications/month}$$

and

$$\text{Average inventory } I = 200 + 150 + 25 = 375 \text{ applications}$$

Thus, we can deduce that

$$\text{Average flow time } T = I/R$$

$$T = 375/1{,}000 \text{ month} = 0.375 \text{ month} = 11.25 \text{ days}$$

Under Process II, therefore, each application spends, on average, 11.25 days with Auto-Moto before an accept/reject decision is made. Compared to the 15 days taken, on average, under Process I, this is a significant reduction.

Another way to reach the same conclusion that $T = 11.25$ days is to do a more detailed analysis and calculate the average flow time of each *type* of application. (Recall that the Initial Review Team at Auto-Moto Financial Services categorizes each application received as Type A, B, or C.) To find the average flow time over *all* applications, we can then take the weighted average of the flow times for each type—in other words, break down Process II into its three subprocesses, initial review, Subgroup A review, and Subgroup B review, and find out how much time applications spend in each of these subprocesses. From that knowledge we can then compute the flow time of each

type of application. The remainder of this section illustrates the detailed computations behind this argument.

As we can see in Figure 3.4, each application starts out in initial review. On average, there are 200 applications with the Initial Review Team. For initial review, the performance measures are denoted with subscript *IR* and are as follows:

$$\text{Throughput } R_{IR} = 1{,}000 \text{ applications/month}$$
$$\text{and}$$
$$\text{Average inventory } I_{IR} = 200 \text{ applications}$$

From this information we can deduce that for initial review,

$$\text{Average flow time } T_{IR} = I_{IR}/R_{IR}$$
$$T_{IR} = 200/1{,}000 \text{ month} = 0.2 \text{ month} = 6 \text{ days}$$

Thus, each application spends, on average, six days in initial review.

Now consider the applications classified as Type A by initial review. Recall that, on average, there are 25 applications with the Subgroup A Review Team. Because 25 percent of all incoming applications are categorized as Type A, on average, 250 of the 1,000 applications received per month are categorized as Type A. We will denote this group with a subscript *A*. So, we have

$$\text{Throughput } R_A = 250 \text{ applications/month}$$
$$\text{Average inventory } I_A = 25 \text{ applications}$$

We can, thus, deduce that

$$\text{Average flow time } T_A = I_A/R_A$$
$$T_A = 25/250 \text{ month} = 0.1 \text{ month} = 3 \text{ days}$$

Type A applications spend, on average, another three days in the process with the Subgroup A Review Team.

Similarly, the Subgroup B Review Team receives 25 percent of incoming applications, or 250 applications per month. It is also given that there are 150 applications with Subgroup B. That is,

$$\text{Throughput } R_B = 250 \text{ applications/month}$$
$$\text{Average inventory } I_B = 150 \text{ applications}$$

We can, thus, deduce that

$$\text{Average flow time } T_B = I_B/R_B$$
$$= 150/250 \text{ month} = 0.6 \text{ month} = 18 \text{ days}$$

Thus, Type B applications spend, on average, another 18 days in the process with the Subgroup B Review Team.

Recall that 50 percent of all incoming applications, or 500 applications per month, are rejected by the Initial Review Team itself. These applications are classified as Type C applications and leave the process immediately. (For sake of consistency, one could say that their additional time spent after IR is $T_C = 0$ so that their inventory $I_C = T_C \times R_C = 0$.)

Recall that the Initial Review Team at Auto-Moto Financial Services categorizes each application received as Type A, B, or C. Each application spends, on average, six days with the Initial Review Team. Type A applications are then reviewed by the Subgroup A Review Team, where they spend an additional three days. Type B applications

are reviewed by the Subgroup B Review Team, where they spend, on average, another 18 days. Type C applications are rejected by the Initial Review Team itself.

Summarizing, we now have computed the average flow time of each *type* of application under Process II:

- Type A applications spend, on average, 9 days in the process.
- Type B applications spend, on average, 24 days in the process.
- Type C applications spend, on average, 6 days in the process.

Finally, we can now find the average flow time across all applications under Process II using this more detailed analysis by taking the weighted average across the three application types. Average flow time across all application types, therefore, is given as follows:

$$T = \frac{R_A}{R_A + R_B + R_C}(T_{IR} + T_A) + \frac{R_B}{R_A + R_B + R_C}(T_{IR} + T_B) + \frac{R_C}{R_A + R_B + R_C}(T_{IR})$$

So,

$$T = \frac{250}{250 + 250 + 500}(6 + 3) + \frac{250}{250 + 250 + 500}(6 + 18) + \frac{500}{250 + 250 + 500}(6)$$

$$T = \frac{250}{1{,}000}(9) + \frac{250}{1{,}000}(24) + \frac{500}{1{,}000}(6) = 11.25 \text{ days}$$

This, indeed, agrees with our earlier (shorter) computation of the average flow time of 11.25 days.

In the analysis so far, we defined flow units according to categories of applications. When evaluating service performance, however, Auto-Moto Financial Services may want to define flow units differently—as applications, approved applications, or rejected applications. Indeed, only approved applications represent customers who provide revenue, and Auto-Moto Financial Services would probably benefit more from reducing their flow time to less than 11.25 days.

Under Process I, the average time spent by an application in the process is 15 days—regardless of whether it is finally approved. Let us now determine how much time the *approved* applications spend with Auto-Moto, under Process II. Under Process II, 70 percent of Type A applications (175 out of 250 per month, on average) are approved, as are 10 percent of Type B applications (25 out of 250 per month, on average). Thus, the aggregate rate at which all applications are approved equals 175 + 25 = 200 applications per month. The average flow time for *approved* applications, denoted by $T_{approved}$, is, again, a weighted average of the flow times of each type of approved application:

Average flow time for approved applications =

$$\begin{aligned} T_{approved} &= \frac{175}{175+25}(T_{IR}+T_A)+\frac{25}{175+25}(T_{IR}+T_B) \\ &= \frac{175}{200}(6+3)+\frac{25}{200}(6+18) \\ &= 10.875 \text{ days} \end{aligned}$$

Similarly, let us now determine the average time an eventually *rejected* application spends with Auto-Moto under Process II, denoted by T_{reject}. Under Process II, 30 percent of Type A applications (75 out of 250 per month, on average) are rejected, as are 90 percent of Type B applications (225 out of 250 per month, on average), as are 100 percent of Type C applications (500 per month, on average).

Average flow time for rejected applications, T_{reject}, is then the weighted average across each of these three types and is given by

$$\begin{aligned} T_{reject} &= \frac{75}{75+225+500}(T_{IR}+T_A)+\frac{225}{75+225+500}(T_{IR}+T_B)+\frac{500}{75+225+500}(T_{IR}) \\ &= \frac{75}{800}(6+3)+\frac{225}{800}(6+18)+\frac{500}{800}(6) \\ &= 11.343 \text{ days} \end{aligned}$$

Process II, therefore, has not only reduced the average overall application flow time but also reduced it *more* for approved customers than for rejected customers. However, 12.5 percent of all approved applications (25 that are categorized as Type B, out of 200 approved each month) spend a lot longer in Process II than in Process I (an average of 24 instead of 15 days). This delay may be a problem for Auto-Moto Financial Services in terms of service performance. Since approved applications represent potential customers, a delay in the approval process may cause some of these applicants to go elsewhere for financing, resulting in a loss of revenue for Auto-Moto.

3.6 ANALYZING FINANCIAL FLOWS THROUGH FINANCIAL STATEMENTS

Our business process-flow paradigm can also be used to analyze financial statements by considering the flow of a financial unit (say, a dollar) through the corporation. Let us return to MBPF Inc. of Example 3.2 and analyze its three financial statements: the firm's income statement, balance sheet, and the more detailed cost of goods sold (COGS) statement for 2011. With an appropriate use of Little's law, this analysis will not only help us understand the current performance of the process but also highlight areas for improvement.

Recall that a key financial measure for any process such as MBPF Inc. is the working capital, which includes the value of process inventories and accounts receivables. The following analysis shows us how to find areas within MBPF Inc. in which a reduction in flow time will result in a significant reduction in inventories and, therefore, the working capital.

In 2011, MBPF operations called for the purchase of both sheet metal (raw materials) and prefabricated bases (purchased parts). Roofs were made in the fabrication area from sheet metal and then assembled with prefabricated bases in the assembly area. Completed garages were stored in the finished goods warehouse until shipped to customers.

In order to conduct our analysis, we need the data contained in the following tables:

- Table 3.4: MBPF's 2011 income statement
- Table 3.5: MBPF's consolidated balance sheet as of December 31, 2011
- Table 3.6: Details concerning process inventories as of December 31, 2011, as well as production costs for 2011

Note that all values in these tables are in millions of dollars and that all data represent end-of-the-year numbers, although we will assume that inventory figures represent average inventory in the process.

3.6.1 Assessing Financial Flow Performance

Our objective is to study cash flows at MBPF in order to determine how long it takes for a cost dollar to be converted into recovered revenue. For that, we need a picture of process-wide cash flows. (Later, to identify more specific areas of improvement within

Table 3.4 MBPF Inc. Consolidated Statements of Income and Retained Earnings for 2011

Net sales	250.0
Costs and expenses	
Cost of goods sold	175.8
Selling, general, and administrative expenses	47.2
Interest expense	4.0
Depreciation	5.6
Other (income) expenses	2.1
Total costs and expenses	234.7
Income before income taxes	15.3
Provision for income taxes	7.0
Net income	8.3
Retained earnings, beginning of year	31.0
Less cash dividends declared	2.1
Retained earnings at end of year	37.2
Net income per common share	0.83
Dividend per common share	0.21

Table 3.5 MBPF Inc. Consolidated Balance Sheet as of December 31, 2011

Current assets	
Cash	2.1
Short-term investments at cost (approximate market)	3.0
Receivables, less allowances of $0.7 million	27.9
Inventories	50.6
Other current assets	4.1
Total current assets	87.7
Property, plant, and equipment (at cost)	
Land	2.1
Buildings	15.3
Machinery and equipment	50.1
Construction in progress	6.7
Subtotal	74.2
Less accumulated depreciation	25.0
Net property, plant, and equipment	49.2
Investments	4.1
Prepaid expenses and other deferred charges	1.9
Other assets	4.0
Total assets	146.9
(Selected) current liabilities	
Payables	11.9

Table 3.6 MBPF Inc. Inventories and Cost of Goods Sold Details

Cost of goods sold	
Raw materials	50.1
Fabrication (L&OH)	60.2
Purchased parts	40.2
Assembly (L&OH)	25.3
Total	175.8
Inventory	
Raw materials (roof)	6.5
Fabrication WIP (roof)	15.1
Purchased parts (base)	8.6
Assembly WIP	10.6
Finished goods	9.8
Total	50.6

the corporation, we will need a more detailed picture.) The flow unit here is a cost dollar, and the process is the entire factory, including the finished-goods warehouse. Incorporating inventory and cash-flow numbers obtained from Table 3.6, a process view of the financial flows through the entire process (factory + finished-goods warehouse) is shown in Figure 3.5. From Table 3.6, we see that raw materials (for roofs) worth \$50.1 million and purchased parts (finished bases) worth \$40.2 million are purchased each year. Labor and overhead costs in roof fabrication total \$60.2 million per year and in final assembly total \$25.3 million per year. Adding all costs, we obtain the annual cost of goods sold, which is \$175.8 million (as shown in Table 3.4). From Table 3.5, we find that inventories at MBPF Inc. total \$50.6 million.

On analyzing the cash flows, we arrive at the following information:

$$\text{Throughput } R = \$\,175.8 \text{ million/year } [\text{Cost of Goods Sold, Table 3.4}]$$
$$\text{Average inventory } I = \$\,50.6 \text{ million } [\text{Inventories, Table 3.5}]$$

Thus, we can deduce average flow time as follows:

$$\begin{aligned}\text{Average flow time } T &= I/R \\ &= 50.6/175.8 \text{ year} \\ &= 0.288 \text{ year} = 14.97 \text{ weeks}\end{aligned}$$

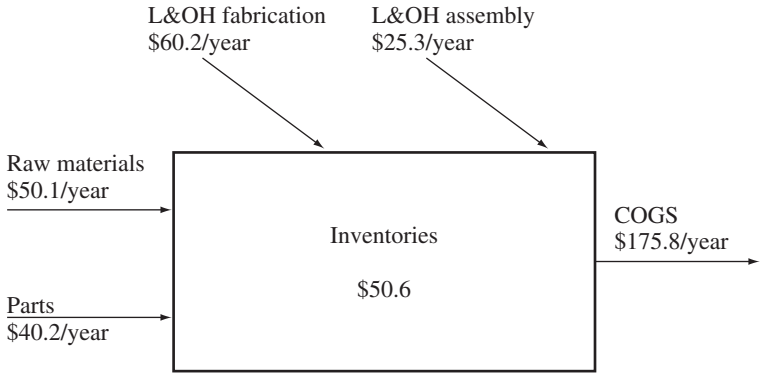


FIGURE 3.5 Financial Flows of MBPF Inc.

FIGURE 3.6 Accounts-Receivable Flows at MBPF Inc.

Alternatively, if we replace the *annual* throughput figure, \$175.8 million per year, by a *weekly* figure, \$3.381 million per week, we can then obtain T in weeks directly as follows:

$$\begin{aligned}\text{Average flow time } T &= I/R \\ &= 50.6/3.381 \text{ weeks} \\ &= 14.97 \text{ weeks}\end{aligned}$$

So, the average dollar invested in the factory spends roughly 15 weeks before it leaves the process through the door of the finished-goods inventory warehouse. In other words, it takes, on average, 14.97 weeks for a dollar invested in the factory to be billed to a customer.

A similar analysis can be performed for the accounts-receivable (AR) department. Let us find out how long it takes, on average, between the time a dollar is billed to a customer and enters AR to the time it is collected as cash from the customer's payment. In this case, process boundaries are defined by the AR department, and the flow unit is a dollar of accounts receivable. From Table 3.4, note that MBPF has annual sales (and thus an annual flow rate through AR) of \$250 million. From Table 3.5, note that accounts receivable in AR total \$27.9 million. Incorporating these numbers, Figure 3.6 presents the process flow view of MBPF's AR department.

When we analyze flows through AR, we arrive at the following information:

$$\begin{aligned}\text{Throughput } R_{AR} &= \$\,250 \text{ million/year } [\text{Net Sales, Table 3.4}] \\ \text{Average inventory } I_{AR} &= \$\,27.9 \text{ million } [\text{Receivables, Table 3.5}]\end{aligned}$$

Accordingly, the average flow time through *AR* (T_{AR}) is

$$\begin{aligned}\text{Average flow time } T_{AR} &= I_{AR}/R_{AR} \\ &= 27.9/250 \text{ year} \\ &= 0.112 \text{ year} = 5.80 \text{ weeks}\end{aligned}$$

In other words, after a sale is made, MBPF must wait, on average, nearly six weeks before sales dollars are collected from the customer.

Finally, the same analysis can be done for the accounts-payable (AP)—or purchasing—process at MBPF Inc. Recall that MBPF purchases both raw materials and parts. Let us find out how long it takes, on average, between the time raw material or parts are received and the supplier bills MBPF (and the bill enters AP) to the time MBPF pays the supplier. In this case, process boundaries are defined by the AP department, and the flow unit is a dollar of accounts payable. From Table 3.6, note that MBPF spends \$50.1 million on raw materials and \$40.2 million on purchased parts per year. The annual flow rate through AP is, therefore, \$50.1 + 40.2 = \$90.3 million. The balance sheet in Table 3.5 shows that the average inventory in purchasing (accounts payables) is

$11.9 million. Letting the subscript AP denote accounts payable, we can use Little's law to determine the average flow time through *AP* department:

$$\begin{aligned} T_{AP} &= I_{AP}/R_{AP} \\ &= 11.9/90.3 \text{ year} \\ &= 0.13 \text{ year} = 6.9 \text{ weeks} \end{aligned}$$

In other words, it takes MBPF, on average, 6.9 weeks to pay a bill.

3.6.2 Cash-to-Cash Cycle Performance

Overall, there is an average lag of about 21 weeks (15 weeks in production and 5.8 weeks in AR) between the point at which cost dollars are invested and the point at which sales dollars are received by MBPF. We call this time of converting cost dollars into sales (21 weeks for MBPF) the *cost-to-cash* cycle for this process. Yet, MBPF only pays for the cost dollars it invests in the form of purchased parts and raw materials after 6.9 weeks. Its total "cash-to-cash" cycle, therefore, is

$$21 - 6.9 = 14.1 \text{ weeks}$$

It is important to realize that flow rates can be expressed in either cost dollars or sales dollars. From Table 3.4, we see that 175.8 million cost dollars result in 250 million in sales dollars. When considering inventories, MBPF must use cost dollars. In contrast, when considering receivables or revenue, MBPF must consider sales dollars. When converting the appropriate rates into flow times, however, all flows are in time units and can be compared.

3.6.3 Targeting Improvement with Detailed Financial Flow Analysis

To identify areas within the process that can benefit most from improvement, we need a more detailed flow analysis. We now consider detailed operations by analyzing dollar flows separately through each of the following areas or departments of the process: raw materials, purchased parts, fabrication, assembly, and finished goods. The flow unit in each case is a cost dollar. A detailed flow diagram is shown in Figure 3.7, with all cost dollar flows in millions of dollars.

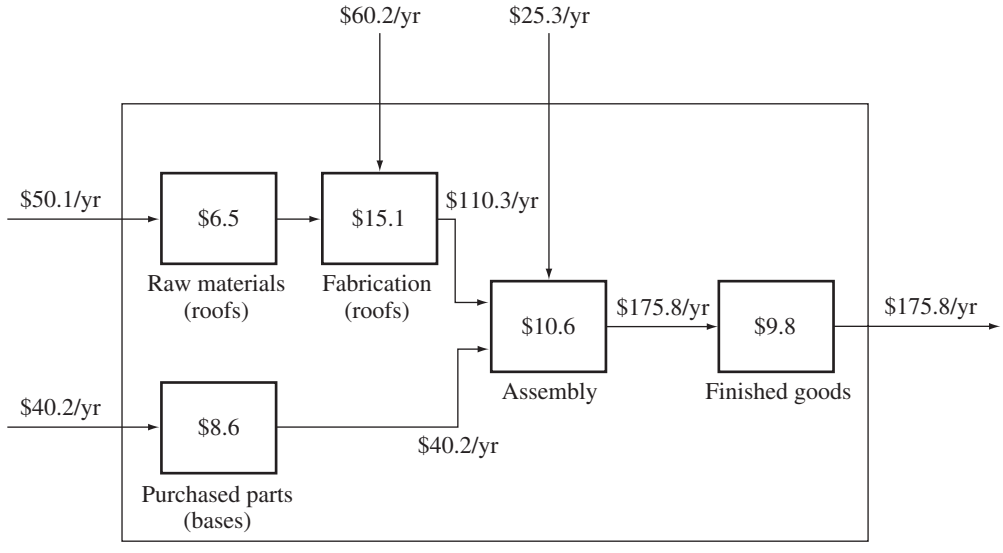


FIGURE 3.7 Detailed Financial Flows at MBPF Inc.

Table 3.7 Flow Times through MBPF Inc.

	Raw Materials	Fabrication	Purchased Parts	Assembly	Finished Goods
Throughput *R*					
$/year	50.1	110.3	40.2	175.8	175.8
$/week*	0.96	2.12	0.77	3.38	3.38
Inventory *I* ($)	6.5	15.1	8.6	10.6	9.8
Flow time $T = I/R$ (weeks)*	6.75	7.12	11.12	3.14	2.90

* Rounding of numbers is done after working through with the initial data.

For each department, we obtain throughput by adding together the cost of inputs and any labor and overhead (L&OH) incurred in the department. So, the throughput rate through fabrication is

$$\begin{aligned} &\$\ 50.1 \text{ million/year in raw materials} \\ +\ &\$\ 60.2 \text{ million in labor and overhead} \\ =\ &\$110.3 \text{ million/year} \end{aligned}$$

The throughput through the assembly area is

$$\begin{aligned} &\$110.3 \text{ million/year in roofs} \\ +\ &\$\ 40.2 \text{ million/year in bases} \\ +\ &\$\ 25.3 \text{ million/year in labor and overhead} \\ =\ &\$175.8 \text{ million/year} \end{aligned}$$

By analyzing the various flows through these four stages, we find the flow times for a cost dollar through each department shown in Table 3.7. (All data originate from Table 3.6.)

Working capital in each department includes the amount of inventory in it. Flow time in each department represents the amount of time a cost dollar spends, on average, in that department. Reducing flow time, therefore, reduces MBPF's required working capital. Knowing this principle, we are prompted to ask: In which department does a reduction of flow time have the greatest impact on working capital? Because inventory equals the product of flow time and throughput, the value of reducing flow time, say, by one week in any department is proportional to its throughput rate. For example, because throughput through the finished-goods warehouse is $3.38 million per week, reducing flow time here by one week saves $3.38 million in working capital (inventory). But because the throughput rate through purchased parts is only $0.77 million per week, a one-week reduction in flow time saves only $0.77 million in working capital. Naturally, the current flow time of an activity represents the maximum potential reduction in flow time.

Both current flow times and the value of reducing them are represented graphically in Figure 3.8. For each department, we plot throughput on the vertical axis and flow time on the horizontal axis. Each department corresponds to a rectangle whose area represents the inventory in the department. Typically, the throughput increases as we go from inflows through the process and end with accounts receivable because it reflects value added.

Observe in Figure 3.8 that a one-week reduction flow time has the largest impact in the AR department because the rectangle for AR represents a flow rate of nearly $5 million per week, which is highest. Thus, reducing the flow time in AR by one week would

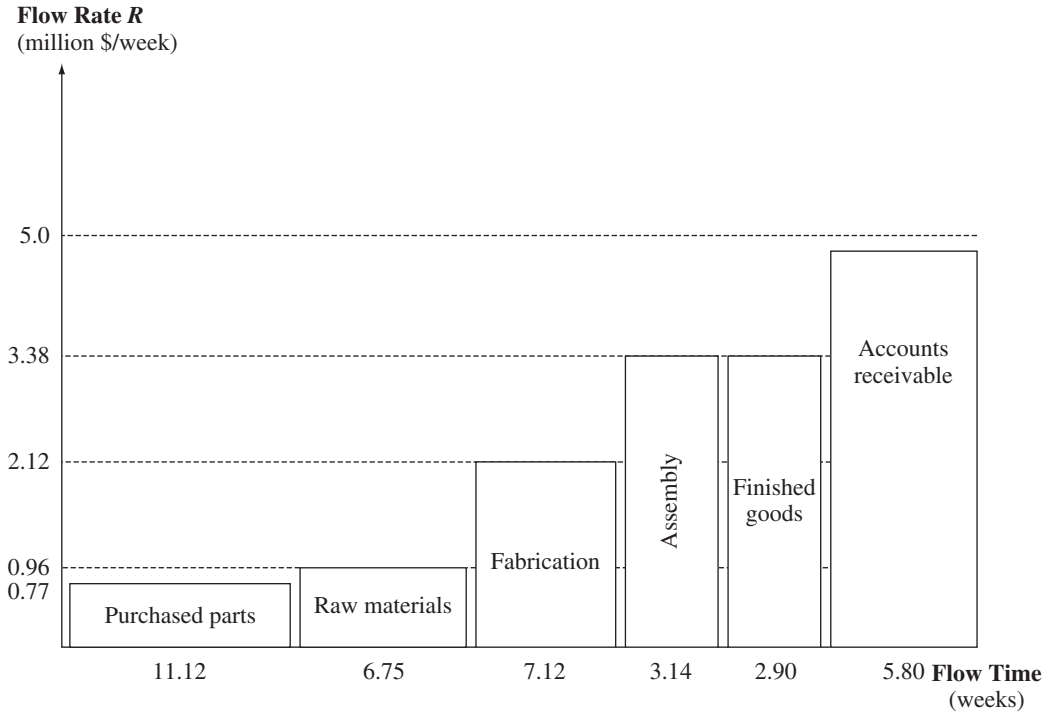


FIGURE 3.8 Representation of Inventory Value at MBPF Inc.

free up nearly $5 million! (Example 3.4 illustrates typical actions to reduce the flow time in AR and free up cash.)

The smallest possible impact of a one-week reduction would be in the purchased parts department; the rectangle in Figure 3.8 that represents it is the shortest and has a flow rate of only $0.77 million. With a flow time of 11.12 weeks, however, the purchased parts department offers the greatest potential to decrease flow time itself.

EXAMPLE 3.4

A portfolio company that provides custodial and security services to business customers needs to reduce its working capital to improve liquidity. The management team is focused on day-to-day operations and has not yet made significant progress in reducing accounts receivable (AR).

One of the board members of the firm, Jeb Bentley, an engineer and former operating manager in the automotive and industrial products industries, suggests applying flow concepts to help management create significant reductions in AR. This will free up about $10 million in cash to pay down debt and increase the value of the firm's investment.

Bentley suggests the following actions:

- The firm will draw a very basic process diagram outlining the length of time that cash is tied up in each stage of the collection process, enabling easier identification of target areas for improvement.
- Inventory of outstanding receivables will be reduced by decreasing the flow time of sending bills. Currently, typically between $2 million and $3 million in bills are in the mail at any given time. By using e-mail to send bills, this inventory will be cut by about 75 percent.
- A policy of ensuring quality at each point in the process ("quality at source") will be implemented to decrease processing time and avoid unnecessary delays. At

present, billings to clients are sent by branch offices to headquarters for review to ensure that they are error free. By pushing this responsibility back onto branch offices, both billing errors and review time will be decreased, as the branches better know their typical billings and the reviewer at headquarters will no longer be a bottleneck. This will result in a further reduction of inventory.

3.7 TWO RELATED PROCESS MEASURES: TAKT TIME AND INVENTORY TURNS (TURNOVER RATIO)

In addition to the average level of inventory, throughput, and the average flow time, practicing operations managers also use two related process measures: takt time and inventory turns.

3.7.1 Takt Time

Takt time is *the reciprocal of throughput and denotes the maximal time that each process resource can devote to a flow unit to keep up with the demand*.

$$\text{Takt time} = 1/R \qquad \textbf{(Equation 3.4)}$$

The word "takt" is German for rhythm or beat. Just like a conductor sets the rhythm for the orchestra, takt time sets the pace of the process to stay synchronized with the demand.

For example, the throughput of 600 passengers per hour of the Vancouver Airport Security Checkpoint implies a takt time of 1 hour per 600 passengers = 0.1 minute per passenger = 6 seconds per passenger. This means that each resource must be able to process a passenger within 6 seconds. Indeed, the Xray scanner only needs to devote 5 seconds per passenger, so that the checkpoint is able to keep up with the demand.

Takt time is a key concept behind lean operations (which we will further discuss in Chapter 10): it translates customer demand into synchronized process design and execution. For example, consider the design of the assembly process of a medium volume passenger car. If

Demand = 150,000 cars per year

Total available production time = 2 × 8 hrs./day × 250 days/yr. = 4,000 hrs./year

Then, Takt time = 4,000 hrs./150,000 cars = 1hr/37.5 cars = 96 secs./car

Obviously, this does not mean that an entire car is assembled in 96 seconds! Rather, it means that the entire assembly operation must be broken down in many steps or "stations," each of which should not require more than 96 seconds on average, potentially with buffers in between. It is exactly this "specialization and division of work" what Scottish economist Adam Smith advocated and what generates the high productivity and throughput of modern processes.

3.7.2 Inventory Turns

Operations managers, accountants, and financial analysts often use the concept of inventory turns or turnover ratio to show how many times the inventory is sold and replaced during a specific period. In the accounting literature, inventory turns is defined as the cost of goods sold divided by average inventory. The cost of goods sold during a given period is nothing other than throughput, expressed in monetary units. Therefore, in our broader view of inventory, **inventory turns**, or **turnover ratio**, is defined as *the ratio of throughput to average inventory*. It is expressed as

$$\text{Inventory turns} = R/I \qquad \textbf{(Equation 3.5)}$$

But we can use Little's law, $I = R \times T$, to come up with an equivalent definition of inventory turns as follows:

$$\begin{aligned} \text{Inventory turns} &= R/I && \text{(by definition)} \\ &= R/(R \times T) && \text{(use Little's law)} \\ &= 1/T && (R \text{ cancels out}) \end{aligned}$$

In other words, inventory turns is the reciprocal of average flow time and thus is a direct operational measure. This directly shows why high turns are attractive: a company with high inventory turns has small flow times and thus is quicker at turning its inputs into sold outputs.

To derive a meaningful turnover ratio, we must specify the flow unit and measure inventory and throughput in the same units. Some organizations measure turns as the ratio of sales to inventory. This measure has a drawback in that sales (a measure of throughput) are expressed in sales dollars but inventory is measured in cost dollars. A better way to calculate turns is the ratio of cost of goods sold (COGS)—labor, materials, and overhead expenses allocated to the products in question—to inventory because both are measured in cost dollars. Example 3.5 illustrates this calculation. Measuring turns as the ratio of sales to inventory can lead to erroneous conclusions when measuring process performance.

EXAMPLE 3.5

Let us return to the MBPF financial statements in Tables 3.4 and 3.5 to analyze inventory turns. We will use cost dollar as the flow unit and designate the factory and the finished-goods warehouse as the process:

$$\begin{aligned} \text{Turns} &= \text{Throughput/Inventory} \\ &= (\$175.8/\text{year})/\$50.6 = 3.47/\text{year} \end{aligned}$$

In other words, during one year MBPF Inc. sells and thus replenishes its average inventory about three and a half times.

3.8 LINKING OPERATIONAL TO FINANCIAL METRICS: VALUING AN IMPROVEMENT

Thus far, we have defined three operational process-performance measures: flow rate, flow time, and inventory. (Recall that takt time and inventory turns are reciprocals of throughput and average flow time respectively.) We have also seen how each can be evaluated for a variety of business process flows. Because Little's law relates these measures through one equation, we can manage only two of them independently; the third measure is then automatically determined. Now let us relate these operational measures to financial measures of process performance. Our goal is to determine when a process change generates an improvement from both operational and financial perspectives.

3.8.1 Linking Operational Improvements to NPV

Net Present Value The financial performance of any business process may be measured by the **net present value (NPV)** of its current and future cash flows. *NPV is a measure of expected aggregate monetary gain or loss that is computed by discounting all expected future cash inflows and outflows to their present value.* Given a sequence of cash flows over a period of future time, a firm's NPV is equivalent to a single present sum such that any risk-neutral investor who is in a position to choose between accepting a future sequence

of cash flows on the one hand or the single sum today values both the same. NPV is calculated by adjusting future cash flows by a discount factor to reflect the time value of money—that is, the principle that a dollar you hold today will be worth more than a dollar you expect to receive in the future. (The discount factor can also be adjusted to account for the investor's risk preferences but we will focus on the time value of money.) The discount factor is based on **rate of return** (r): *the reward that an investor demands for accepting payment delayed by one period of time*.

Let C_t represent the cash flow in period t, starting from period $t = 0$ and ending at period $t = n$. The NPV of these cash flows is found by first discounting each cash flow C_t—that is, multiplying it by the discount factor of $1/(1 + r)^t$—and then summing all those discounted cash flows:

$$\text{NPV} = C_o + \sum_{t=1}^{n} \frac{C_t}{(1 + r)^t}$$

(Net present value can also be directly computed using built-in spreadsheet functions, such as "NPV" in Microsoft Excel.)

Sales Volume and Cash Flows The true throughput for any business process is measured by sales volume—the number of units sold. If MBPF Inc. produces 2,200 garages per week while market demand is for 1,000 garages per week, the throughput as measured by sales would be 1,000 garages per week, while the remaining 1,200 garages per week would simply build up as inventory. If, however, production is only 800 garages per week and demand is 1,000 per week, then finished-goods inventory will soon be depleted, after which actual sales—and thus throughput—will be only 800 garages per week. True throughput, as measured by long-term average sales rate, therefore, is the minimum of its output and market demand. Note that positive cash flows result from the revenue received from product sales (there may be other revenue sources, but product sales will be a major contributor). Thus, we can assume that positive cash flows are correlated with throughput. An increase in throughput (sales) thus increases positive cash flows.

Negative cash flows typically result from investment in resources, interest expense, and operating expense (labor + overhead). Interest expense is correlated with the amount of inventory in the process. Reducing inventory in the process reduces the company's working capital requirement. In turn, this reduction lowers its interest expense, thereby reducing negative cash flows. As we noted in Chapter 1, we define process cost as the total cost incurred in producing and delivering outputs. Negative cash flows, therefore, can also be reduced by reducing process cost because negative cash flows decrease when the process entails lower cost to produce the outputs.

Now we can assess when a change in the process can be called an improvement. From the financial perspective, a change is an improvement *if and only if* it increases NPV. A change may increase both positive and negative cash flows, which is an improvement only if the NPV of the increase in positive cash flows exceeds the NPV of the increase in negative cash flows. A change can certainly be called an improvement if it either increases positive cash flows without increasing negative cash flows or decreases negative cash flows without decreasing positive cash flows. This situation is equivalent to addressing the following three questions:

1. Has true process throughput (as measured by sales) risen without any increase in inventories or process cost?
2. Has process inventory declined without any reduction in throughput or increase in process cost?
3. Has process cost declined without any reduction in throughput or increase in inventory?

The first two questions really ask whether flow time has been reduced without any increase in process cost. All three questions are quite similar to those raised by E. M. Goldratt in his efforts to identify and characterize process improvement (Goldratt, 1992). All three questions address only simple instances in which we know that NPV will go up because of the change. In more complicated scenarios in which both positive and negative cash flows change, we must evaluate NPV before characterizing a change as an improvement.

3.8.2 Linking Operational Improvements to Financial Ratios

In addition to NPV, it is also informative to consider the impact of operations on financial ratios. Here, we follow the approach by Chopra and Meindl (2009). The definitions of financial measures in this section are taken from Dyckman, Magee, and Pfeiffer (2011).

From a shareholder perspective, return on equity (ROE) is the main summary measure of a firm's performance.

$$ROE = \frac{Net\ Income}{Average\ Shareholder\ Equity}$$

Whereas ROE measures the return on investment made by a firm's shareholders, return on assets (ROA) measures the return earned on each dollar invested by the firm in assets.

$$\text{ROA} = \frac{\text{Earnings before interest}}{\text{Average Total Assets}} = \frac{\text{Net Income} + \left[\text{Interest expense} \times (1 - \text{tax rate})\right]}{\text{Average Total Assets}}$$

Consider Amazon.com's financial performance shown in Table 3.8. In 2009, Amazon achieved ROE = 17.2 percent (902/5257) and ROA = 6.7 percent [902+34*(1−.35)]/13813). The difference between ROE and ROA is referred to as return on financial leverage (ROFL). In 2009, Amazon had ROFL = 10.5 percent. ROFL captures the amount of ROE that can be attributed to financial leverage (accounts payable, debt etc.). In Amazon's case, a significant portion of the financial leverage in 2009 came from accounts payable rather than debt. Thus, an important ratio that defines financial leverage is accounts payable turnover (APT).[1]

$$APT = \frac{Cost\ of\ goods\ sold}{Accounts\ payable}$$

In Amazon's case, in 2009 APT = 2.58. A small APT indicates that Amazon was able to use the money it owed suppliers to finance a considerable fraction of its operations. Amazon effectively financed its own operations for about 52/2.58 = 20.18 weeks with its suppliers' money.

ROA can be written as the product of two ratios—profit margin and asset turnover—as shown in the following equation:

$$\begin{aligned}\text{ROA} &= \frac{\text{Earnings before interest}}{\text{Sales Revenue}} \times \frac{\text{Sales Revenue}}{\text{Total Assets}} \\ &= \text{Profit Margin} \times \text{Assets Turnover}\end{aligned}$$

Thus, a firm can increase ROA by growing the profit margin and/or increasing the asset turnover. In 2009, Amazon achieved a profit margin of 3.8 percent [902+34*(1−.35)]/24509). Profit margin can be improved by getting better prices or by reducing the

[1]Ideally the numerator in APT should be cost of purchased materials and not cost of goods sold. However, public companies in the United States do not report their cost of purchased materials and, therefore, COGS is often used as a proxy.

Table 3.8 Selected Financial Data for Amazon.com Inc.

Year ended December 31 ($ millions)	2009	2008
Net operating revenues	24,509	19,166
Cost of goods sold	18,978	14,896
Gross profit	5,531	4,270
Selling, General, and Administrative expense	4,402	3,428
Operating income	1,129	842
Interest expense	34	71
Other income (loss) – net	66	130
Income before income taxes	1,161	901
Income taxes	253	247
Net income	902	645
Assets		
Cash and cash equivalents	3,444	2,769
Short term investments	2,922	958
Net receivables	1,260	1,031
Inventories	2,171	1,399
Total current assets	9,797	6,157
Property, plant, and equipment	1,290	854
Goodwill	1,234	438
Other assets	1,492	705
Total assets	13,813	8,314
Liabilities and Stockholder Equity		
Accounts payable	7,364	4,687
Short term debt		59
Total current liability	7,364	4,746
Long term debt	109	533
Other liabilities	1,083	363
Total liabilities	8,556	5,642
Stockholder equity	5,257	2,672

various expenses incurred. A responsive operation can allow a firm to provide high value to a customer, thus potentially getting higher prices. Good operations management can also allow a firm to decrease the expenses incurred to serve customer demand. In Amazon's case, a very significant expense is outbound shipping cost. In its 2009 annual report, the company reported outbound shipping costs of $1.77 billion. After accounting for shipping revenue, the net loss on outbound shipping was reported to be $849 million, about the same order of magnitude as net income. Clearly, a reduction in outbound shipping costs can have a significant impact on Amazon's profit margin.

The key components of asset turnover are accounts receivable turnover (ART), inventory turnover (INVT), and property, plant and equipment turnover (PPET). These are defined as follows:

$$ART = \frac{Sales\ revenue}{Accounts\ receivable}; INVT = \frac{Cost\ of\ goods\ sold}{Inventories}; PPET = \frac{Sales\ revenue}{PP\ \&\ E}$$

Amazon achieved accounts receivable turnover of 19.45 per year in 2009, which indicates that it collected its money from sales relatively quickly (in about 52/19.45 = 2.7 weeks,

on average, after making a sale). Amazon turned its inventory about 8.74 times in 2009 and had PPET = 19. Thus, inventory sat with Amazon for about 52/8.74 = 5.95 weeks on average, and each dollar invested in property, plant, and equipment supported about $19 of sales in 2009. Following our earlier discussion, this is equivalent to a cash-to-cash cycle of 5.95 + 2.7 − 20.18 = − 11.53 weeks.

Amazon can improve its asset turnover by turning its inventory quicker or using its existing warehousing infrastructure to support a higher level of sales (or decreasing the warehousing infrastructure needed to support the existing level of sales).

From our brief discussion of Amazon's financial statements it is clear that operations management activities such as transportation, inventory, and warehousing have a significant impact on financial performance. In general, an affirmative answer to any of the three questions posed previously also means there is an increase in **return on total assets**.

In this chapter, we have established a relationship between three key operational measures and some common financial measures. Our discussion indicates that improvements in the three key operational measures translate into improvements in financial measures as well. Therefore, the operational measures of throughput, inventory, and flow time are leading indicators of financial performance.

Summary

The first chapter in this book discussed the importance of identifying operational measures that are good leading indicators of customer satisfaction and the financial performance of a process. This chapter introduces three key operational measures that characterize the flow of units through the process: throughput, inventory, and flow time. Throughput is the rate at which units flow through the process. Inventory is the number of flow units within the process boundaries at a given point in time. Flow time is the time it takes for a specific flow unit to be transformed from input to output. The three operational measures can be applied to processes with a variety of flow units, including money, customers, data, material, and orders.

For a stable process, the three operational measures are related through Little's law, which states that average inventory is the product of average throughput and average flow time. In other words, managers need to track and control only two of the three measures—average throughput and average inventory, typically, which then determine average flow time.

These operational measures are leading indicators of financial performance. Inventory is a measure of tied-up capital (for manufacturing) or customers who are waiting (for services). For a manufacturing firm, a decrease in inventory indicates a drop in working capital requirements. Throughput measures the rate at which the output of the process is being sold. An increase in throughput indicates increased revenues and also increased profits if the product has positive margin. A higher throughput means a smaller takt time and thus less available time for each resource to process a flow unit and keep up with demand. Flow time measures how long it takes to transform orders and invested cash into products. A faster flow time means higher inventory turns and relatively lower working capital requirements. An improvement in the three operational measures thus leads to an improvement in long-term financial measures, such as net present value and return on investment.

Key Equations and Symbols

(Equation 3.1) $\Delta R(t) = R_i(t) - R_o(t)$
(Equation 3.2) $I(t_2) - I(t_1) = \Delta R \times (t_2 - t_1)$
(Equation 3.3) Little's law: $I = R \times T$
(Equation 3.4) Takt time = 1/R
(Equation 3.5) Inventory turns = R/I

where
$R_i(t)$: Instantaneous inflow rate
$R_o(t)$: Instantaneous outflow rate
$\Delta R(t)$: Instantaneous inventory accumulation (or buildup) rate
$I(t)$: Inventory at time t
I: Average inventory
R: Throughput or average flow rate
T: Average flow time

Key Terms

- Average flow rate
- Average flow time
- Flow rate
- Flow time
- Instantaneous inventory accumulation (buildup) rate
- Inventory buildup diagram
- Takt time
- Inventory turns
- Little's law
- Net present value (NPV)
- Process flow measures
- Rate of return
- Return on total assets
- Stable process
- Throughput
- Turnover ratio

Discussion Questions

3.1 Why is it important to look at aggregate flow performance, as measured by average inventory, average flow time, and average throughput?

3.2 Discuss why it is often easier to measure average inventory and average throughput rather than average flow time.

3.3 How can a manager determine the minimal set of operational measures that should be tracked on a daily basis to predict the financial performances of a process?

3.4 The Internal Revenue's Department of Tax Regulations writes regulations in accordance with laws passed by Congress. On average, the department completes 300 projects per year. The *Wall Street Journal* reports that, as of October 11, 2011, the number of projects currently "on the department's plate" is 588. Nevertheless, the department head claims that average time to complete a project is less than six months. Do you have any reason to disagree? Why or why not?

3.5 The *Wall Street Journal* reported that "although GM and Toyota are operating with the same number of inventory turns, Toyota's throughput is twice that of GM." The discrepancy, concluded the writer, "could be due to much faster flow times and lower inventories by virtue of Toyota's production system." With which of the following deductions do you agree?

- **a.** The two statements are consistent.
- **b.** The two statements are inconsistent: If both have the same inventory turns, they have the same flow time; but Toyota has higher average inventory than GM.
- **c.** The two statements are inconsistent: If both have the same inventory turns, they have the same flow time; but Toyota has lower average inventory than GM.
- **d.** The two statements are inconsistent: If both have the same inventory turns, they have the same average inventory; but Toyota has higher flow time than GM.
- **e.** The two statements are inconsistent: if both have the same inventory turns, they have the same average inventory; but Toyota has lower flow time than GM.

3.6 Is there a difference between low inventories and fast inventory turnover?

3.7 Is there a difference between flow time and takt time? Illustrate with an example.

3.8 Why is it preferable to have a short cost-to-cash cycle, and how can that be achieved?

Exercises

***3.1** A bank finds that the average number of people waiting in line during lunch hour is 10. On average, during this period, 2 people per minute leave the bank after receiving service. On average, how long do bank customers wait in line?

3.2 At the drive-through counter of a fast-food outlet, an average of 10 cars waits in line. The manager wants to determine if the length of the line is having any impact on potential sales. A study reveals that, on average, 2 cars per minute try to enter the drive-through area, but 25 percent of the drivers of these cars are dismayed by the long line and simply move on without placing orders. Assume that no car that enters the line leaves without service. On average, how long does a car spend in the drive-through line?

3.3 Checking accounts at a local bank carry an average balance of $3,000. The bank turns over its balance 6 times a year. On average, how many dollars flow through the bank each month?

***3.4** A hospital emergency room (ER) is currently organized so that all patients register through an initial check-in process. At his or her turn, each patient is seen by a doctor and then exits the process, either with a prescription or with admission to the hospital. Currently, 55 people per hour arrive at the ER, 10 percent of who are admitted to the hospital. On

average, 7 people are waiting to be registered and 34 are registered and waiting to see a doctor. The registration process takes, on average, 2 minutes per patient. Among patients who receive prescriptions, average time spent with a doctor is 5 minutes. Among those admitted to the hospital, average time is 30 minutes. On average, how long does a patient spend in the ER? On average, how many patients are being examined by doctors? On average, how many patients are there in the ER? Assume the process to be stable; that is, average inflow rate equals average outflow rate.

3.5 A triage system has been proposed for the ER described in Exercise 3.4. Under the proposed triage plan, entering patients will be registered as before. They will then be quickly examined by a nurse practitioner who will classify them as Simple Prescriptions or Potential Admits. While Simple Prescriptions will move on to an area staffed for regular care, Potential Admits will be taken to the emergency area. Planners anticipate that the initial examination will take 3 minutes. They expect that, on average, 20 patients will be waiting to register and 5 will be waiting to be seen by the triage nurse. Recall that registration takes an average of 2 minutes per patient. The triage nurse is expected to take an average of 1 minute per patient. Planners expect the Simple Prescriptions area to have, on average, 15 patients waiting to be seen. As before, once a patient's turn comes, each will take 5 minutes of a doctor's time. The hospital anticipates that, on average, the emergency area will have only 1 patient waiting to be seen. As before, once that patient's turn comes, he or she will take 30 minutes of a doctor's time. Assume that, as before, 90 percent of all patients are Simple Prescriptions. Assume, too, that the triage nurse is 100 percent accurate in making classifications. Under the proposed plan, how long, on average, will a patient spend in the ER? On average, how long will a Potential Admit spend in the ER? On average, how many patients will be in the ER? Assume the process to be stable; that is, average inflow rate equals average outflow rate.

3.6 Refer to Exercise 3.5. Once the triage system is put in place, it performs quite close to expectations. All data conform to planners' expectations except for one set—the classifications made by the nurse practitioner. Assume that the triage nurse has been sending 91 percent of all patients to the Simple Prescription area when in fact only 90 percent should have been so classified. The remaining 1 percent is discovered when transferred to the emergency area by a doctor. Assume all other information from Exercise 3.5 to be valid. On average, how long does a patient spend in the ER? On average, how long does a Potential Admit spend in the ER? On average, how many patients are in the ER? Assume the process to be stable; that is, average inflow rate equals average outflow rate.

3.7 Orange Juice Inc. produces and markets fruit juice. During the orange harvest season, trucks bring oranges from the fields to the processing plant during a workday that runs from 7 A.M. to 6 P.M. On peak days, approximately 10,000 kilograms of oranges are trucked in per hour. Trucks dump their contents in a holding bin with a storage capacity of 6,000 kilograms. When the bin is full, incoming trucks must wait until it has sufficient available space. A conveyor moves oranges from the bins to the processing plant. The plant is configured to deal with an average harvesting day, and maximum throughput (flow rate) is 8,000 kilograms per hour.

Assuming that oranges arrive continuously over time, construct an inventory buildup diagram for Orange Juice Inc. In order to process all the oranges delivered during the day, how long must the plant operate on peak days? (Assume, too, that because Orange Juice Inc. makes fresh juice, it cannot store oranges.) Assuming, finally, that each truck holds about 1,000 kilograms of oranges, at what point during the day must a truck first wait before unloading into the storage bin? What is the maximum amount of time that a truck must wait? How long will trucks wait, on average? Among trucks that do wait, how long is the average wait?

3.8 Jasper Valley Motors (JVM) is a family-run auto dealership selling both new and used vehicles. In an average month, JVM sells a total of 160 vehicles. New vehicles represent 60 percent of sales, and used vehicles represent 40 percent of sales. Max has recently taken over the business from his father. His father always emphasized the importance of carefully managing the dealership's inventory. Inventory financing was a significant expense for JVM. Max's father consequently taught him to keep inventory turns as high as possible.

- **a.** Examining the dealership's performance over recent years, Max discovered that JVM had been turning its inventory (including both new and used vehicles) at a rate of 8 times per year. What is JVM's average inventory (including both new and used vehicles)?
- **b.** Drilling down into the numbers, Max has determined that the dealership's new and used businesses appear to behave differently. He has determined that turns of new vehicles are 7.2 per year, while turns of used vehicles are 9.6 per year. Holding a new vehicle in inventory for a month costs JVM roughly $175. Holding the average used vehicle in inventory for a month costs roughly $145. What are JVM's average monthly financing costs per vehicle?
- **c.** A consulting firm has suggested that JVM subscribe to its monthly market analysis service. They claim that their program will allow JVM to maintain its current sales rate of new cars while reducing

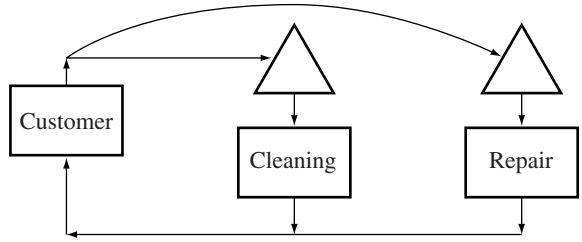


FIGURE 3.9 Flowchart for Cheapest Car Rentals

the amount of time a new car sits in inventory before being sold by 20 percent. Assuming the consulting firm's claim is true, how much should Max be willing to pay for the service?

3.9 Cheapest Car Rental rents cars at the Chicago airport. The car rental market consists of two segments: the short-term segment, which rents for an average of 0.5 week, and the medium-term segment, which rents for an average of 2 weeks. Cheapest currently rents an average of 200 cars a week to the short-term segment and 100 cars a week to the medium-term segment.

Approximately 20 percent of the cars returned (evenly distributed across both segments) are found to be defective and in need of repairs before they can be made available for rent again. The remaining cars not needing repairs are cleaned, filled with gas, and made available for rent. On average, there are 100 cars waiting to be cleaned. The average cost of this operation is $5 per car. Cars needing repairs spend an average of 2 weeks in the repair shop and incur an average cost of $150 per car. Assume that cars are rented as soon as they are available for rent, that is, as soon as they have been cleaned or repaired.

Short-term renters pay $200 per week, while medium-term renters pay $120 per week. The flow of cars is shown in Figure 3.9.

a. Identify throughput, inventory, and flow time at each stage.

b. What profit does Cheapest earn per week with the current system? Assume that each car loses $40 in value per week because of depreciation.

c. Cheapest is comparing two possible improvements:

1. Decrease time in repairs from 2 weeks to 1 week.
2. Decrease cost per repair from $150 per car to $120 per car while keeping flow time in repairs at 2 weeks.

Assume that the effort that is required in each case is the same. Which change do you think will be more effective? Why?

3.10 The Evanstonian is an upscale independent hotel that caters to both business and leisure travelers. On average, one-third of the guests checking in each day are leisure travelers. Leisure travelers generally stay for 3.6 nights—twice as long as the average business customer.

a. On an average day, 135 guests check into The Evanstonian. On average, how many guests of each type are in the hotel on any given day?

b. How many times per month does the hotel turn over its inventory of guests (assume 30 days per month)?

c. The average business traveler pays a rate of $250 per night, while leisure travelers pay an average rate of $210 per night. What is the average revenue The Evanstonian receives per night per occupied room?

3.11 ABC Corporation's consolidated income statement and balance sheet for the years 2011 and 2012 is shown in Table 3.9 (in thousands of dollars).

How do you think cash flow performance in 2011 compares with that of 2012 in the factory as well as accounts receivable? Do you think 2012 is an improvement over 2011? Why?

Table 3.9 Selected Income Statement and Balance Sheet Figures for ABC Corporation

	2011	2012
Net revenues	$99,621	$110,644
Cost of goods sold	$97,380	$98,350
Current assets		
Cash	$13,491	$8,079
Inventories	$20,880	$25,200
Accounts receivable	$21,596	$22,872

Selected Bibliography

Chopra, S., and P. Meindl. *Supply Chain Management: Strategy, Planning, and Operation* 4th ed. Upper Saddle River, NJ: Prentice Hall, 2009.

Dyckman, Thomas R., Robert P. Magee, Glenn M. Pfeiffer. *Financial Accounting*. 3rd edition. Westmont, Il: Cambridge Business Publishers, 2011.

Goldratt, E. M. *The Goal.* 2nd ed. Great Barrington, Mass.: North River Press, 1992.

Hopp, W. J., and M. L. Spearman. *Factory Physics.* Chicago: Irwin, 1996.

Chapter

4

Flow-Time Analysis

INTRODUCTION

Zhang & Associates,[1] a financial advisory branch of American Express, provides comprehensive financial advisory and asset management services to individuals with high net worth. Zhang & Associates' new client process is typical of the industry. It entails a sequence of meetings with the new customer that continues until a mutually acceptable plan of action is identified and implemented. A major weakness of the process is the amount of time required for each customer—a new client with a simple portfolio is processed in four to six weeks; the time for individuals with more complex situations could be much longer.

Recently, the company has redesigned the process so that it can be completed in two weeks. Zhang & Associates' customers are delighted with the faster service. In addition, the company was able to better utilize its resources and improve its relation with existing customers.

How do companies such as Zhang & Associates manage their processes to reduce flow time? We discuss this question in the next few chapters.

In the previous chapter, we introduced three important measures of process performance, namely, flow time, flow rate, and inventory. We also showed how Little's law establishes a fundamental relationship among the averages of these three measures. We

[1]We are grateful to Ms. Lynn L. Chen-Zhang for bringing this example to our attention.

stressed the importance of these measures and applied Little's law to a macro-level performance evaluation of a business process. In this chapter and in Chapter 5, we lay the foundation for more detailed process analysis. Our goal is to understand the factors that affect the three key performance measures and the levers that can be manipulated to improve process performance.

We begin with the concept of flow time. Recall from Chapter 3 that the flow time of a given flow unit is the total amount of time required by that unit to flow through the process from entry to exit. For any given process, the flow time of different flow units varies substantially. *The average flow time of the individual flow units* is called the **flow time of a process**.

Process flow time is a valuable measure of process performance for several reasons:

1. Flow time affects delivery-response time, a key product attribute that customers value, as discussed in Chapter 1. The less time customers must wait for a good or a service, the greater the value for the customer. Also, the shorter the delivery-response time, the more quickly a firm can collect revenues, thus improving its financial performance.
2. Short flow times in the production and delivery process reduce the inventory (by Little's law) and associated costs.
3. A shorter flow time in a firm's new product development process enables the firm to more quickly introduce the product to the market, which is a major source of competitive advantage. Likewise, it enables the firm to bring more generations of a product to market within a given amount of time.
4. In markets featuring short product life cycles, shorter manufacturing-process flow times allow firms to delay production closer to the time of sale and thus gain valuable market information, avoid product obsolescence, and minimize the inventory required.
5. Short flow times result in fast feedback and correction of quality problems, as we see in Chapters 9 and 10.
6. Finally, flow time is important because it is an integrative measure of overall process performance—short flow time frequently requires a high level of overall operational excellence. For instance, a process rife with quality problems would typically display the longer flow times required for inspections, rework (fixing defective products so that they conform to specifications), and so forth.

In this chapter, we study the factors that determine process flow time and examine some ways to manage and reduce it. In Section 4.1, we discuss how flow time can be measured. In Section 4.2, we examine how a process could be presented graphically in the form of a flowchart. In Section 4.3, we examine how the process flow time can be determined from the flowchart by identifying the critical path. In Section 4.4, we examine the roles of **activity time** and waiting time as they relate to total flow time, and introduce the concepts of theoretical flow time and of flow-time efficiency. In Section 4.5, we identify some key managerial levers for managing flow time, with particular emphasis on reducing theoretical flow time.

4.1 FLOW-TIME MEASUREMENT

The flow time of a given process can be determined in two independent ways: (i) by direct observation and (ii) by application of Little's law (Chapter 3). A direct measurement can be made as follows:

1. Observe the process over a specified, extended period of time.
2. Select a random sample of flow units over the specified period.
3. Measure the flow time, from entry to exit, of each flow unit in the sample.
4. Compute the average of the flow times measured.

To use the indirect approach, we measure the throughput, R, and average inventory, I, and then compute average flow time T by using Little's law:

$$I = R \times T$$

The process for measuring R is described in Chapter 5. Average inventory could be measured as follows:

1. Observe the process over a specified, extended period of time.
2. Select a random sample of points of time during the specified period.
3. Measure the actual inventory within the system boundaries at each point of time in the sample.
4. Compute the average of the inventory values measured.

EXAMPLE 4.1

Consider the process of credit approval at a small bank. To estimate the flow time of the process using the *direct* approach, a sample of 50 applications was taken. For each application selected, the flow time was observed. An average of the 50 observations was found to be 20.85 working days. This is a direct estimate of the process flow time.

To estimate the flow time using the indirect approach, a sample of 10 days was selected during a given period, and for each day selected, the number of applications within the process was counted. The average of these 10 observations was found to be 215. This is an estimate of the average inventory in the system. Also, the throughput for this period was determined (see Chapter 5), to be 10 applications per day. The flow time of the process is then given by

$$T = I/R = 215/10 = 21.5 \text{ days}$$

This provides us with an alternative estimate of the process flow time. Naturally, we expect the two estimates to be close, but not identical, because of the randomness of sampling.

Both approaches outlined in the previous section require that flow units, as well as the entry and exit points to the process, be carefully specified. Consider, for example, the process of baking bread. Depending on the specific purpose of the analysis, flow units can be taken to be loaves of bread, pounds of flour, cost or revenue dollars, and so forth. Similarly, there are many reasonable choices for the entry point such as the time the flour is purchased or the time that the mixing operation commences. Finally, the exit point could be taken, for instance, at the point when a baked loaf is unloaded from the oven or when the bread is shipped out of the bakery. Clearly, many other reasonable options are possible, depending on the nature of the analysis. The thing to note is that each possible selection will result in a different definition—and a different numeric value—of the throughput, flow time, and inventory. Some of these choices are demonstrated in Example 4.2:

EXAMPLE 4.2

The research department of Golden Touch Securities is charged with releasing research reports for the benefit of Golden's customers and brokers. The firm's reports can be classified into two types: new releases, which require significant investment of research effort, and updates, which are much smaller in scope. In a Typical month the department releases 20 new releases, for a combined value (sales price) of $40,000, and 40

updates, for a combined value of $20,000. On average, there are 10 open (unfinished) new releases and 8 open updates.

If we are interested in the flow time of reports, irrespective of type, we can define the flow unit of the process as one report of either type. Under this definition the throughput of the process is 20 + 40 = 60 reports per month, and the Inventory is 10 + 8 = 18 reports. The process flow time is

$$T = I/R = 18/60 = 0.3 \text{ month}$$

On the other hand, if we are interested in following the flow of (sales) dollars through the system, we can define the flow unit as a sales-dollar. The throughput under this convention is 40,000 + 20,000 = $60,000 per month, and the inventory is $10 \times 2{,}000 + 8 \times 500 = \$24{,}000$. In this case,

$$T = I/R = 24{,}000/60{,}000 = 0.4 \text{ month}$$

Why the difference? In the first case, the average gives equal weight to each report, independent of type. In the second case, the average gives equal weight to each dollar, and thus gives higher weight to new releases, which are more expensive and take longer to complete. We emphasize that each such set of specifications is correct, and we are free to select the one that best matches the objectives of the study, in this case operational versus financial. However, once the specifications are made, we must define inventory, throughput, and flow time consistently.

4.2 THE PROCESS FLOWCHART

In Chapter 1, we described a process as having five elements, namely, inputs and outputs, flow units, a network of activities and buffers, resources allocated to activities, and an information structure. A **process flowchart** is *a graphical representation of the network structure of the process*. Process flowcharts were originally developed to coordinate large projects involving complex sets of activities and considerable resources. Over the years, however, flowcharts have been found useful for analyzing, managing, and documenting almost any business process. Naturally, several types of graphical representations have been developed as various additional aspects of the process have been included.

The most elementary form of the process depicts

- Activities as rectangles.
- Precedence relationships as solid arrows.
- Events (such as the beginning or end of the process) as ovals.

EXAMPLE 4.3

To illustrate the function of a simple process flowchart, consider the manufacturing process at Wonder Shed Inc., a manufacturer of storage sheds. The manufacturing process involves the procurement of sheets of steel that will be used to form both the roof and the base of each shed.

The first step involves separating the material needed for the roof from that needed for the base. Then the roof and the base can be fabricated in parallel, or simultaneously. Roof fabrication involves first punching then forming the roof to shape. Base fabrication entails the punching-and-forming process plus a subassembly operation. Fabricated roofs and bases are then assembled into finished sheds that are subsequently inspected for quality assurance. A list of activities needed to fabricate a roof, fabricate a base, and assemble a shed is given in Table 4.1.

A flowchart of the process is shown in Figure 4.1.

Table 4.1 Activity List for Wonder Shed Inc.

	Activity
1	Separate the roof and base materials
2	Punch the base
3	Punch the roof
4	Form the base
5	Form the roof
6	Subassemble the base
7	Assemble
8	Inspect

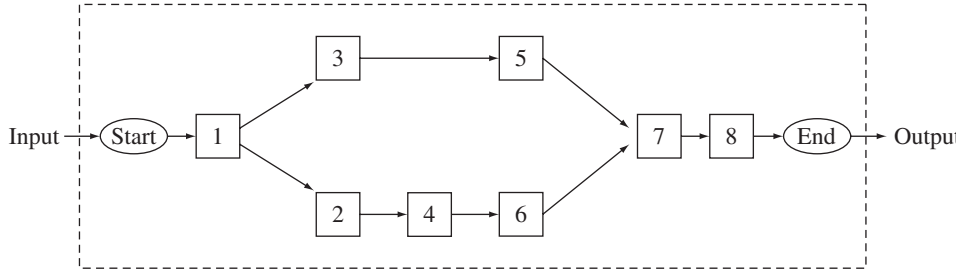


FIGURE 4.1 Process Flowchart for Wonder Shed Inc.

Various additional aspects of the process are sometimes added to enrich the basic chart. For instance, buffers, which are designated locations within the process for the accumulation of inventory, are typically represented as triangles. Decisions, which are activities at which flow is "routed" into several continuing routes, resulting in splitting of the flows, are sometimes depicted for the purpose of emphasis as diamonds. The roles of the various resources can be emphasized by partitioning the flowchart into several horizontally distinct bands, one for each resource. Similarly, information flows can be distinguished from physical flows by the use of dashed arrows.

What is the appropriate level of detail for a given flowchart? Obviously, this depends on the nature of the required analysis. In many cases, it is useful to consider a given process at various levels of detail. In Appendix 4.1, we examine how this could be achieved using a technique called cascading.

4.3 FLOW TIME AND CRITICAL PATHS

In the previous section, we examined how the network structure of the process can be represented as a simple flowchart diagram. In this section we show how the flow time of the process can be computed by combining the flowchart with information about the time to complete the various activities represented on it.

The flow time of a process is the amount of time required by an average flow unit to flow through the process from entry to exit. Similarly, the **flow time of a given activity** within the process is *the time required by an average flow unit to flow through the activity*. This could be measured by observing the specific activity over an extended time interval, or estimated based on experience. Note that the flow time of the process, and of each of its activities, consists of periods of activity interspersed with periods of waiting. We expand on this point in Section 4.4.

The flow times of the various activities in the process, coupled with the sequence in the various activities performed, allow us to compute the flow time of the entire process. To see how this can be done, consider, first, a simple process in which all activities are carried out sequentially, one following the other. In this case, the process flowchart consists of a single path (or route) of activities connecting the entry and exit points of the process. Because all activities along this path must be completed sequentially, the total time required to complete the process equals the sum of the individual activity times along the path. For instance, if a process consists of three activities, A, B, and C, which must be performed sequentially, and if each activity requires 10 minutes, then the total time required to process a flow unit is the sum of the three activity times, that is, $10 + 10 + 10 = 30$ minutes.

However, most processes consist of a combination of sequential and parallel activities, resulting in a process chart that contains several paths running from start to finish. For each path, the flow time along that path is the sum of the flow times of the activities that constitute the path. Now, a flow unit can exit the process only after all the activities along *all* the paths are completed. The flow time of the process, therefore, must equal the flow time of *the longest path in the process flowchart*—the **critical path**. *Activities that lie on a critical path* are called **critical activities**. We illustrate these concepts in Example 4.4:

EXAMPLE 4.4

To demonstrate the computation of flow time of a process, consider again the Wonder Shed illustration begun in Example 4.1. Table 4.2 complements the information in Table 4.1 by adding the flow time of each activity.

Note that our flowchart in Figure 4.1 shows two paths connecting the beginning and end of the process:

$$\text{Path 1 (roof): Start} \rightarrow 1 \rightarrow 3 \rightarrow 5 \rightarrow 7 \rightarrow 8 \rightarrow \text{End}$$
$$\text{Path 2 (base): Start} \rightarrow 1 \rightarrow 2 \rightarrow 4 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow \text{End}$$

The flow time of Path 1, followed by the roofs, is 120 minutes:

	Activity	Flow Time
1	Separate	20
3	Punch the roof	25
5	Form the roof	20
7	Assemble	15
8	Inspect	40
	Total	120 minutes

Table 4.2 Flow Times at Wonder Shed Inc. (in minutes)

	Activity	Flow Time (minutes)
1	Separate	20
2	Punch the base	35
3	Punch the roof	25
4	Form the base	10
5	Form the roof	20
6	Subassemble the base	30
7	Assemble	15
8	Inspect	40

The flow time of Path 2, followed by the bases, is 150 minutes:

	Activity	Flow Time
1	Separate	20
2	Punch the base	35
4	Form the base	10
6	Subassemble the base	30
7	Assemble	15
8	Inspect	40
	Total	150 minutes

Thus, the flow time of the process is 150 minutes, and Path 2 (making the bases) is the critical path.

For simple processes, the critical path can often be determined as in Example 4.4, by computing the flow time of each path. For complex processes, however, there may be too many paths. Consequently, we may sometimes need a more efficient approach to identify the critical path. This is provided in Appendix 4.2.

The critical activities of a process are extremely important for managing flow time since they determine the flow time of the entire process: A delay in completing any critical activity results directly in a corresponding delay in processing the flow unit. As a result, management of the critical path is of paramount significance. In contrast, activities that are not critical can be delayed, to a degree, without affecting the flow time. Thus, they require a reduced level of monitoring by management.

In some situations, flow units need to be repeated at some activities due, for example, to the effects of defects. The role of rework and its effect on the flow time of the process is examined in Appendix 4.3.

4.4 THEORETICAL FLOW TIME AND THE ROLE OF WAITING

As a flow unit travels through the various activities which make up the process, it undergoes periods of waiting interspersed with periods of activity. Thus, for each activity, we can break down the flow time of the activity into its waiting and activity components:

$$\text{Flow time} = \text{Activity time} + \text{Waiting time} \qquad \textbf{(Equation 4.1)}$$

The **theoretical flow time** of a process is the *minimum amount of time required for a flow unit to flow through the process from entry to exit, without any waiting or interruptions.* It can be computed from the flowchart of the process using the same approach as for computing the flow time, by using data on activity time instead of flow time:

EXAMPLE 4.5

Let us return to the Wonder Shed Inc. example. Table 4.3 list the activity time for each activity:

Computing the activity times along paths 1 and 2 we get

$$\text{Path 1: } = 5 + 10 + 5 + 10 + 15 = 45 \text{ minutes}$$
$$\text{Path 2: } = 5 + 15 + 5 + 10 + 10 + 15 = 60 \text{ minutes}$$

Table 4.3 Activity Times at Wonder Shed Inc. (in minutes)

	Activity	Activity Time (minutes)
1	Separate	5
2	Punch the base	15
3	Punch the roof	10
4	Form the base	5
5	Form the roof	5
6	Subassemble the base	10
7	Assemble	10
8	Inspect	15

Thus the theoretical flow time for the process is 60 minutes. The difference between the theoretical flow time of the process (60 minutes) and the flow time (150 minutes) is due to the effects of waiting.

In this example, the critical path for the flow time, Path 2, is the same as for the theoretical flow time (compare Examples 4.4 and 4.5). In general, the amount of waiting along different paths may vary. Thus, when activity time replaces flow time, the relative lengths of the various paths—and the identity of the critical path—may change.

4.4.1 Flow-Time Efficiency

By comparing the average flow time with the theoretical value, we can get an indication of the relative fraction of the flow time that is caused by waiting, as opposed to activity. We formalize this observation using the concept of **flow-time efficiency**—*the ratio between theoretical flow time and the flow time of a given process*. Formally:

$$\text{Flow-time efficiency} = \text{Theoretical flow time}/\text{Average flow time} \quad \textbf{(Equation 4.2)}$$

For example, for the case of Wonder Shed Inc., the flow time is 150 minutes (Example 4.4) while the theoretical flow time is only 60 minutes (Example 4.5). Thus,

$$\text{Flow-time efficiency} = 60/150 = 40\%$$

The values of the flow-time efficiency for a variety of processes were studied by Blackburn (1992) and are excerpted in Table 4.4. Their surprisingly low values underscore the significance of reducing waiting time as we try to improve flow-time performance.

Table 4.4 Flow-Time Efficiency of Business Processes

Industry	Process	Flow Time	Theoretical Flow Time	Flow-Time Efficiency
Life insurance	New policy application	72 hours	7 minutes	0.16%
Consumer packaging	New graphic design	18 days	2 hours	0.14%
Commercial bank	Consumer loan	24 hours	34 minutes	2.36%
Hospital	Patient billing	10 days	3 hours	3.75%
Auto manufacture	Financial closing	11 days	5 hours	5.68%

The theoretical flow time of the process, which represents the total activity time required to process a flow unit, can itself be broken down into two components as follows:

Theoretical flow time = Value-adding flow time + Non-value-adding flow time

Reducing non-value-adding flow time is often a powerful way to save time and money. This topic is discussed in detail in Section 4.5.2.

The following example summarizes several of the points covered in this chapter:

EXAMPLE 4.6

Valley of Hope Hospital has been under recent pressure from stakeholders to improve cost efficiency and customer service. In response, the hospital has undertaken a series of process-improvement initiatives. One of the first processes targeted for improvement was the X-ray service. A major concern identified by both physicians and patients has been the amount of time required to obtain an X-ray. In addition, management would like to make sure that available resources are utilized efficiently.

A process-improvement team was set up to study the X-ray service process and recommend improvements. The team identified the point of entry into the process as the instant that a patient leaves the physician's office to walk to the X-ray lab. The point of exit was defined as the instant that both the patient and the completed X-ray film are ready to enter the physician's office for diagnosis. The unit of flow is a patient.

To determine the flow time of the existing process, a random sample of 50 patients was observed over a two-week period. For each patient, the team recorded times of entry and exit from the X-ray service process. The difference between these two times was then used as a measure of flow time for each patient. The average of the 50 data points was 154 minutes. This figure, then, serves as an estimate of the average flow time for the X-ray service process.

To further study process flow time, the team examined the entire process in detail and broke it down into the constituent activities identified in Table 4.5 as value-added (VA) or non-value-added (NVA).

The corresponding process flowchart is shown in Figure 4.2. It depicts all activities and the precedence relationships among them. For example, Activity 2 must be completed before Activity 3 can begin. Meanwhile, Activity 1 can be carried out simultaneously with Activities 2 and 3. Note that the classification of activities to VA and NVA is somewhat subjective, and may depend on the specific details of the situation:

The team analyzing the process flowchart identified four activity paths:

Path 1: Start $\rightarrow$ 1 $\rightarrow$ 4 $\rightarrow$ 5 $\rightarrow$ 6 $\rightarrow$ 7 $\rightarrow$ 8 $\rightarrow$ 9 $\rightarrow$ 10 $\rightarrow$ End

Path 2: Start $\rightarrow$ 2 $\rightarrow$ 3 $\rightarrow$ 4 $\rightarrow$ 5 $\rightarrow$ 6 $\rightarrow$ 7 $\rightarrow$ 8 $\rightarrow$ 9 $\rightarrow$ 10 $\rightarrow$ End

Path 3: Start $\rightarrow$ 1 $\rightarrow$ 4 $\rightarrow$ 5 $\rightarrow$ 6 $\rightarrow$ 7 $\rightarrow$ 8 $\rightarrow$ 11 $\rightarrow$ End

Path 4: Start $\rightarrow$ 2 $\rightarrow$ 3 $\rightarrow$ 4 $\rightarrow$ 5 $\rightarrow$ 6 $\rightarrow$ 7 $\rightarrow$ 8 $\rightarrow$ 11 $\rightarrow$ End

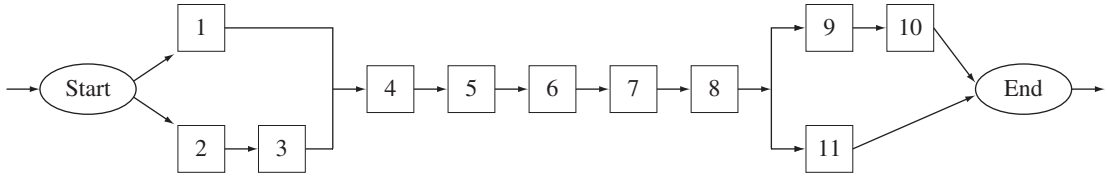


FIGURE 4.2 Flowchart for the X-Ray-Service Process at Valley of Hope Hospital

Table 4.5 The X-Ray-Service Process at Valley of Hope Hospital

Activity/Event	Description	Type
Start	Patient leaves the physician's office.	
1	Patient walks to the X-ray lab.	NVA
2	The X-ray request travels to the X-ray lab by a messenger.	NVA
3	An X-ray technician fills out a standard form based on the information supplied by the physician.	NVA
4	The receptionist receives from the patient information concerning insurance, prepares and signs a claim form, and sends to the insurer.	NVA
5	Patient undresses in preparation for X-ray.	NVA
6	A lab technician takes X-rays.	VA
7	A darkroom technician develops X-rays.	VA
8	A lab technician prepares X-rays for transfer.	NVA
9	Patient puts on clothes and gets ready to leave lab.	NVA
10	Patient walks back to the physician's office.	NVA
11	The X-rays are transferred to the physician by a messenger.	NVA
End	Patient and X-rays arrive at the physician's office.	

Next, another sample of 50 patients was studied over a two-week period. For each patient, the activity time required to perform each activity was recorded. These are listed in Table 4.6:

The theoretical flow time along these four paths are

$$\text{Path } 1 = 50 \text{ minutes}$$
$$\text{Path } 2 = 69 \text{ minutes}$$
$$\text{Path } 3 = 60 \text{ minutes}$$
$$\text{Path } 4 = 79 \text{ minutes}$$

Table 4.6 Work Content in X-Ray-Service Process Activities

Activity	Activity Time (minutes)
Start	—
1	7
2	20
3	6
4	5
5	3
6	7.5
7	15
8	2.5
9	3
10	7
11	20
End	—

Path 4, therefore, is the critical path, yielding a theoretical flow time of the process as 79 minutes.

What is the flow-time efficiency of the process?

$$\text{Flow-time efficiency} = \text{Theoretical flow time}/\text{Average flow time} = 79/154 = 51\%$$

This means that waiting corresponds to roughly half the time in this process. Obviously, the challenge this poses to the management of Valley of Hope Hospital is whether some of this waiting can be eliminated. Also, note that of the 79 minutes of theoretical flow time, the only activities which are value adding are Activities 6 and 7. Thus the value adding time of the process is 7.5 + 15 = 22.5 minutes, which is less than 15 percent of the average flow time. Indeed, Valley of Hope has ample opportunities to improve the process!

4.5 LEVERS FOR MANAGING THEORETICAL FLOW TIME

How can managers reduce the flow time of a process? As we have seen, the only way to reduce the flow time is to shorten the length of every critical path. We have also seen that the flow time is the sum of two components—waiting time and activity time. Because these two components arise from different sources, the levers available for managing each are naturally distinct. The main levers for reducing waiting time in a process are:

(i) Managing the effects of congestion (Chapters 5 and 8)
(ii) Reducing batch sizes (Chapter 6)
(iii) Reducing safety buffers (Chapter 7)
(iv) Synchronizing flows (Chapter 10)

In this section, however, we examine the levers available for managing the activity part of the flow time—the theoretical flow time.

There are five basic approaches to shortening a critical path:

(i) Move work content off the critical path ("work in parallel")
(ii) Eliminate non-value-adding activities ("work smarter")
(iii) Reduce the amount of rework ("do it right the first time")
(iv) Modify the product mix ("do the quickest things first")
(v) Increase the speed of operation ("work faster")

There are significant differences between these five approaches. The first approach is one of *restructuring:* It leaves the total amount of work per unit unaffected, but manages the sequencing of the various activities in order to reduce the length of the critical path. The second approach is one of *elimination.* It leaves the network structure of the process as is, but reduces the total amount of work required for activities along the critical path. The third approach depends on setting a robust *quality management system* (Chapter 9). The fourth approach is one of *prioritization.* It gives priority to flow units that can be processed faster—to the extent allowed by the market. The fifth approach relies on working at a *faster* rate. Naturally, for each specific situation, the relative merits of these five approaches will vary.

It is critical to remember, that whatever approach we take, it must be directed towards the critical path: Reducing the work content of noncritical activities does *not* reduce the theoretical flow time. However, such reduction may still be useful for other reasons, such as decreasing total processing costs, increasing process capacity (see Chapter 5), and reducing the potential for errors and defects. In the following section, we examine each of these approaches more fully.

4.5.1 Moving Work Off the Critical Path

One of the best ways to reduce the theoretical flow time is by moving work off the critical path and into paths that do not affect process flow time. This task can be accomplished in one of two ways:

1. Move work off the critical path to a noncritical activity.
2. Move work off the critical path to the "outer loop" (pre- or postprocessing).

In either case, the work must still be done, but the critical path is shortened.

Moving work from a critical to a noncritical path means redesigning the process so that critical activities are performed in parallel rather than sequentially. Consider, for example, the conventional approach to software development, which consists of five steps in sequence: specification, design, development, documentation, and testing. Clearly, testing and documentation can be carried out in parallel. Moreover, it is often not necessary to complete the development of the software in order to start preparing the user manual. Thus, it is possible to perform some aspects of software design, development, testing, and documentation in parallel.

For another example, consider the contemporary practice of concurrent engineering. Traditionally, activities such as product design, process planning, and prototyping are performed sequentially. By modifying the process to increase parallelism we can speed the process considerably. (We describe the applications of concurrent engineering more fully in Chapter 10.)

Moving activities to the so-called outer loop means performing them either before the countdown for the process starts or after it ends, as defined by the process boundary, an approach that is also called pre- or postprocessing. For example, in the case of the hospital admission process, it is often possible to accomplish work such as verifying insurance, preparing and signing consent forms, and listing of allergies even before the patient shows up at the hospital. As another example, consider the process of changing a "make-to-order" production system into a "make-to-stock" system. Instead of assembling a complete hamburger after receiving a customer order, it may be possible to precook beef patties and keep them ready prior to the lunchtime rush. As far as customer flow is concerned, theoretical flow time will be reduced because the production of the beef patty has been moved to the outer loop of the "order-fulfillment process." Note, however, that because it produces units prior to demand, this strategy affects the "material flow process" in the opposite fashion. In case of hamburgers, of course, it may also affect taste and quality.

4.5.2 Reduce Non-Value-Adding Activities

It is a common observation that some of the work done by individuals and organizations is not essential for the product or service produced. The idea that such nonessential work should be systematically eliminated—saving time and money—can be traced to the scientific management approach used by Frederic Taylor (1911), Frank Gilbreth (1911), and their followers, as discussed in Chapter 2. Originally, the approach was used for optimizing work by individual workers, typically in manual tasks, such as laying bricks, loading coal, or typing a manuscript. However, the core ideas of this approach are still valid today in the much broader context of a general business process, both in service and in manufacturing.

Value-adding activities *are those activities that increase the economic value of a flow unit from the perspective of the customer (that is the customer values such activities, and is willing to pay for them).* Performing surgery, flying an airplane, serving meals in a restaurant, manufacturing an item in a factory, and dispensing a loan by a bank are examples of activities which are typically value-adding.

Non-value-adding activities *are activities that do not directly increase the value of a flow unit.* For example moving work or workers among various locations, setting up machines, scheduling activities or personnel, sorting, storing, counting, filling out forms, participating in meetings, obtaining approvals or maintaining equipment are typically non-value-adding.

Non-value-adding activities come in two types: (i) Non-value-adding work that is necessary to support the current process and (ii) Non-value-adding work that does not. Obviously, non-value-adding activities of the second type should be eliminated outright. However, activities of the first type can also be eliminated if the process is redesigned. For example, a process that is rife with high fractions of defectives may require a sorting station to separate the defective from the good units. The sorting activity is a non-value-adding activity, but is necessary given the process. However, if the process capability is increased so that no defectives are produced (see Chapter 10), the sorting activity becomes unnecessary, and, therefore, one that could be eliminated. As another example, consider the accounts-payable process. The primary value-adding activity of this process is paying the bills in an accurate and timely fashion. However, the accounts-payable department typically spends much of its time performing other activities, such as reconciling contradictory information, verifying, matching documents, and investigating discrepancies. Such activities do not add value but are still necessary, given the process utilized. They can be eliminated, however, if the process is modified. Hammer and Champy (1993), for instance, report that the accounts-payable department at Ford was reengineered to eliminate unnecessary steps with a dramatic reduction in flow time and cost. One of the innovations introduced was the elimination of issuing and processing invoices and the rejection of any shipment that does not conform exactly to the purchase order. For details, see Hammer and Champy (1993).

4.5.3 Reduce the Amount of Rework

Decreasing the amount of repeat work can often be achieved by process-improvement techniques such as statistical process control, design for manufacturability, process foolproofing, and workforce training (see Chapter 10). In data-rich environments, the key principle is to strive toward a process that "touches" any particular data input just once since the common custom of entering the same data over and over again adds time (as well as cost and errors). The effect of rework on flow time is explored in Appendix 4.3.

4.5.4 Modifying the Product Mix

Most processes involve a mix of products, characterized by different flow times for the various units of flow. If we give priority to flow units that move through the process faster, the overall flow time of the process will decrease. Of course, product mix is often dictated by the market, and even when the organization has some control over it, there may be other relevant factors, such as profitability, resource-utilization issues, and market considerations. Nevertheless, modifying the mix and serving more customers or jobs that could be handled faster is sometimes an effective way to reduce average flow time. We elaborate more on the product mix decision in Chapter 5.

4.4.5 Increase the Speed of Operations

The speed at which an activity is performed can be improved by acquiring faster equipment, increasing allocated resources, or offering incentives for faster work. Such steps often require either financial investment in faster equipment or modified incentives for labor resources. Consider, for instance, a manual checkout counter at a local grocery store. The speed of this operation can be increased by any of the following methods:

using bar codes with a scanner, adding a second worker to bag products, or instituting proper incentives, coupled with training and better equipment so that checkout personnel work faster, without increasing error rates or jeopardizing service quality. In a research-and-development laboratory, the so-called dedicated teams that concentrate fully on one activity rather than working on several projects simultaneously can increase the speed at which a particular research activity is carried out.

4.4.6 Zhang & Associates Revisited

We close with a more detailed description of the process improvement activities undertaken by Zhang & Associates introduced earlier in the chapter. As mentioned in the introduction, the company provides comprehensive financial advisory and asset management services to high-net-worth individuals. The company has redesigned its new client process and cut the flow time from six to four weeks and in some cases, considerably more, to two weeks. We now review how Zhang & Associates was able to achieve these results, utilizing some of the levers mentioned in Section 4.4.

THE OLD PROCESS

The point of entry into the old process was when a new client arrived to meet the adviser for the introductory meeting. During the meeting, the adviser took notes about the client's financial information and listed details such as the client's stocks, life insurance policies, and bank accounts. After the meeting, the adviser reviewed the notes with a staff member of the planning department, called a "paraplanner." The paraplanner typed the information into financial planning software and prepared a general financial plan. At this stage, the paraplanner often found that the adviser had neglected to obtain all the relevant information during the first meeting. In such cases, the paraplanner contacted the adviser, who in turn contacted the client to obtain the necessary information.

Some clients had advanced planning needs, such as estate planning, in which case the completed general financial plan was forwarded to the advanced planning department. The professionals in the advanced planning department, often attorneys or certified public accountants, reviewed the general financial plan and discussed the client situation with the adviser before providing recommendations. After the financial plan was completed, the adviser conducted a second meeting with the client to go over the plan. The client took the plan home for detailed review. If the plan was acceptable to the client, a third meeting was scheduled to finalize the plan and sign the necessary documents. If the client was not satisfied with the plan, another cycle of consultations with the staff of the advanced planning department was initiated. The process was completed when the plan was approved by the customer and finally implemented.

The process typically required one month to one-and-a-half months for completion. The time could be substantially longer if the client's situation was complicated, requiring a fourth or even a fifth meeting.

THE NEW PROCESS

Zhang & Associates has recently implemented a new process. The key differences from the old process can be summarized as follows.

A "homework" package is sent to the client and is completed by the client before the first meeting. This set of forms reveals critical personal and financial information. Clients can obtain assistance to complete the forms by calling the adviser's office.

The first meeting involves everyone required for devising the financial plan, including the adviser, the paraplanner, and all the relevant advanced planning professionals. By the end of the meeting, everyone understands all the issues involved and will be able to work on their parts of the plan simultaneously.

The paraplanner receives the various plans from the various participants and assembles them into a comprehensive plan. The competed plan is mailed to the client for reviewing prior to the second meeting. The customer reviews the plan and can resolve any remaining issues before the meeting, thereby significantly reducing the possible need for a third or fourth meeting.

In preparation for the second meeting, the paraplanner prepares all the forms that are needed for the implementation of the plan. During that meeting, the adviser and the paraplanner go over the plan with the client and address all remaining questions or concerns. In most cases, the client approves the plan at this point and signs the necessary forms.

NEW PROCESS SUCCESS

By adopting the new process, Zhang & Associates was able to reduce the time for completion of the process to two weeks! The improvement is achieved by a combination of the following levers:

1. Move work off the critical path:
   a. Move work to the outer loop (premeeting "homework").
   b. Work in parallel (after the first meeting, all professionals can work simultaneously since they are all on board).
2. Elimination of non-value-adding work (only two meetings instead of three or more).

The results of improving the process are quite dramatic: Clients are happy because they don't have to wait up to two months for their plan to be implemented. In addition, the company utilizes its personnel more effectively and saves costs because the advisers focus on activities that add value, such as investment portfolio design. Another bonus is that the new process builds a relationship of confidence between the customer and the paraplanner (in the old process, the customer never met the paraplanner). This increases customer satisfaction. Zhang & Associates further capitalized on this relationship in order to improve its meetings with the existing customers along the same lines.

The new process imposes additional demands on the professional staff, especially the paraplanners and advanced planning staff. In the old process, the adviser was the only one meeting the customer, and all the other staff members were kept "behind the scenes." In contrast, in the new process, everyone is in the "front line."

Summary

In the preceding chapters, we introduced three important measures of process performance—flow time, flow rate, and inventory—and discussed the relationship among them (Little's law). In this chapter, we focus on flow time, the first of these measures. Reducing the flow time of a given process is important to the organization and its customers since it increases responsiveness, customer satisfaction, and financial performance. In addition, reducing flow time often requires improvements in other aspects of the process, such as reduction in defects and rework, leading to improved quality.

Flow time of a given process can be measured either directly, by observing the time taken by various flow units, or indirectly, by observing the throughput and inventory and then utilizing Little's law.

The process flowchart is a graphical representation of a process that breaks it down into a set of activities and identifies their interrelationships. Using data on the flow time of the individual activities, the flow time of the process can be determined as the flow time along the critical path.

The flow time of the process can be broken down to two components: activity and waiting. The

theoretical flow time measures the activity time: It represents the flow time of a unit through the process without any waiting. Flow-time efficiency is a metric that gives an indication of the extent of waiting in a process. The theoretical flow time can be further decomposed into its value-adding and non-value-adding components.

To improve the flow time of a process, we need to focus on the activities along the critical path. Flow time can be reduced by affecting either of its two components: activity time or waiting time. Levers for reducing waiting time include managing congestion, reducing batch sizes, reducing safety buffers, and synchronizing flows. In this chapter, we have focused on the theoretical flow time.

There are five key managerial levers for reducing theoretical flow time. First, theoretical flow time can be reduced by moving some work content off the critical path. Two ways to achieve this include working in parallel rather than in sequence, thereby moving work to a noncritical activity, and moving work to the outer loop. Second, theoretical flow time can be reduced by elimination of non-value-adding aspects of the activity. As some non-value-adding activity may be necessary for the proper functioning of the process, this may require a redesign of the process. Third, activity time could be shortened by reducing the amount of rework. Fourth, theoretical flow time can be altered by selecting a suitable product mix. Finally, one can increase the speed at which the activity is performed.

Key Equations and Symbols

(Equation 4.1) Flow time = Activity time + Waiting time

(Equation 4.2) Flow-time efficiency = Theoretical flow time / Average flow time

Key Terms

- Activity time
- Critical activities
- Critical path
- Flow time of a given activity
- Flow time of a process
- Flow-time efficiency
- Forward scheduling
- Non-value-adding activities
- Process flowchart
- Slack time
- Subprocess
- Theoretical flow time
- Total flow time
- Value-adding activities

Discussion Questions

4.1 Examine a service encounter such as a visit to a restaurant. Give a rough estimate of the average flow time, the theoretical flow time, and the flow-time efficiency.

4.2 Provide an example from your work environment of how flow time could be improved using the following levers:
- **a.** "Work smarter"
- **b.** "Do it right the first time"
- **c.** "Work faster"

4.3 Describe the process used by your bank for approving a home mortgage. In particular, highlight the activities that are done or could be done in parallel. What are the pros and cons of doing these activities in parallel?

4.4 A group of MBA students is preparing a business case for submission as a final project for their operations management course. Describe the process used by the group from the perspective of working in parallel versus working sequentially.

4.5 The speed at which pit crews replace flat tires at car races such as the Indy 500 is amazing. Discuss the effects of "moving work to the outer loop" in this context.

4.6 How long does it take your company to process a business expense form? What is your estimate of the theoretical time required for this process? What is the flow-time efficiency?

4.7 How long does it take your company to process a customer complaint? Draw a process map of the process utilized, and discuss how some of the levers used in this chapter can be used to speed up this process.

4.8 Decreasing the activity time of an activity always improves the flow time. Comment.

Table 4.7 The Traffic Court

Defendant	Arrival	Departure	Time with Judge (minutes)	Time Paying Fine (minutes)
1	8:45	9:30	1	5
2	8:45	9:45	1.5	2
3	8:45	12:05	2	3
4	8:50	12:55	1.5	5
5	8:50	10:35	1	2
6	8:55	9:20	1	0
7	8:55	11:35	2	2
8	9:00	10:45	3	0
9	9:00	12:55	1	2
10	9:00	9:20	1.5	3

Exercises

***4.1** The Traffic Court of King James County operates between the hours of 9 A.M. and 1 P.M. Each morning, roughly at 9 p.m., 200 defendants show up for trial involving traffic violations, such as speeding, illegal parking, and ignoring a stop sign. On Monday, June 10, 2003, a sample of 10 defendants was selected at random by a consultant. For each defendant, the consultant recorded the actual time spent in discussion with the judge and the time paying the fine (not including waiting). Also recorded were the times the defendant arrived and left the court. The data are summarized in Table 4.7.

a. Estimate the flow time of the process.
b. Estimate the theoretical flow time of the process.
c. What is the flow-time efficiency?

4.2 Wonder Shed Inc. (Example 4.3) produces, in addition to the standard model, a deluxe version for the discriminating customer. The production process for the two models is identical and is depicted in Figure 4.1. The activity times for deluxe models is listed in Table 4.8. All the times mentioned represent flow time at the various activities, and include the effects of waiting.

a. Compute the process flow time for producing a deluxe shed.
b. What is the impact on flow time of the process if the flow time of Activity 2 is increased to 40 minutes?
c. What is the impact on flow time of the process if the flow time of Activity 3 is reduced to 40 minutes?

4.3 The Evanstonian is an upscale independent hotel that caters to both business and leisure travelers. When a guest calls room service at The Evanstonian, the room-service manager takes down the order. The service manager then submits an order ticket to the kitchen to begin preparing the food. She also gives an order to the sommelier (i.e., the wine waiter) to fetch wine from the cellar and to prepare any other alcoholic beverages. Finally, she assigns the order to a waiter.

Table 4.8 Flow time Deluxe Model, Wonder Shed Inc.

Activity	Flow time Deluxe (minutes)
1	20
2	35
3	45
4	10
5	45
6	30
7	25
8	40

It takes 4 minutes to take down the order and to assign the work to the kitchen, sommelier, and waiter. It takes the kitchen 18 minutes to prepare the typical order. It takes the sommelier 6 minutes to prepare the drinks for the order. While the kitchen and the sommelier are doing their tasks, the waiter readies a cart (i.e., puts a tablecloth on the cart and gathers silverware). This takes 10 minutes per order.

Once the food, wine, and cart are ready, the waiter delivers it to the guest's room. It takes the waiter 12 minutes to deliver the meal to the customer. It takes the waiter additional 4 minutes to return to the station and debit the guest's account. All the times mentioned represent flow time at the various activities, and include the effects of waiting.

a. Draw a process map for the room-service process: from receipt of order to delivery of food
b. What is the flow time of the process?
c. What is the effect on the process flow time if the waiter could prepare the cart in 8 minutes, instead of 10?
d. What is the effect on the process flow time if the waiter could deliver the order in 10 minutes, instead of 12?
e. Now redefine the process to begin upon receipt of order, and end upon debit of account. Repeat parts a and b

4.4 A home insurance application consists of two forms: F1, which relates to the home owner, and F2, which relates to the property. On receipt, each application is processed, recorded, and separated into F1 and F2. This operation requires 10 minutes. F1 requires Activity A for 15 minutes per unit and then Activity B for 10 minutes per unit. F2 requires Activity C for 20 minutes per unit. F1 and F2 are then combined and further processed by a loan officer for 15 minutes. All the times mentioned represent flow time at the various activities, and include the effects of waiting.
a. Draw a process flowchart for the processes.
b. What is the flow time?
c. What is the effect on flow time if 50 percent of F1 forms must repeat Activity A one more time due to quality problems? (See Appendix 4.3.)

***4.5** The Vancouver International Airport Authority, described in Chapter 3, manages and operates the Vancouver International Airport (YVR). Its focus on safety, security, and customer service has contributed to YVR's ranking among the top 10 airports in the world. To maintain its excellent customer service standards and in anticipation of new government regulations, airport management sought to take leadership in improving customer flow through its airport security checkpoints.

To understand flow, management started with a single security line comprising an X-ray scanner for carry on items and a screening station for passengers. Arriving customers first prepare themselves for the inspection by removing belts, coats and shoes, emptying their pockets, and separating electronic gear from other personal items. They then deposit all bags in trays on the scanner and proceed personally to the screening station. Once the screening is completed, passengers retrieve their belongings, put on their shoes, belts, and coats, and exit the facility.

On average, it takes passengers 30 seconds to prepare for the line, and to place all carry-on items in the trays for the X-ray scanner. The X-ray scanner takes 40 seconds per tray, and the average passenger utilizes 1.5 trays. The personal screening station requires 30 seconds per person. Finally, retrieving of belongings and getting reorganized takes 60 seconds. All the times mentioned represent activity time at the various activities and do not include the effects of waiting.
a. Draw a process map for the security check process.
b. What is theoretical flow time of the security check process?
c. A sample of 20 passengers was selected at random, and the time required for each to clear the security check was recorded. The average of the individual times was 530 seconds. What is the process flow-time efficiency?
d. What is the impact on theoretical flow time of the process if the personal screening activity is expedited to 20 seconds?

Selected Bibliography

Blackburn, J. D. "Time-Based Competition: White-Collar Activities." *Business Horizons* 35, no. 4 (1992): 96–101.

Chase, R. B., N. J. Aquilano, and F. R. Jacobs. *Production and Operations Management.* 10th ed. Chicago: Irwin McGraw-Hill, 2004.

Eppen, G. D., F. Gould, C. Schmidt, J. Moore, and L. Weatherford. *Introductory Management Science.* 5th ed. Upper Saddle River, N.J.: Prentice Hall, 1998.

Evans, J. R. *Applied Production and Operations Management.* 4th ed. Minneapolis: West Publishing, 1994.

Gilbreth, F. B. *Motions Study.* New York: Van Nostrand, 1911.

Hammer, M., and J. Champy. *Reengineering the Corporation.* New York: HarperBusiness, 1993.

Bohn, R. E. *Kristen's Cookie Company.* Harvard Business School Case 9-686-093. Cambridge, Mass.: Harvard Business School, 1986.

Kaniegel, R. *The One Best Way: Fredrick Winslow Taylor and the Enigma of Efficiency.* New York: Penguin, 1997.

Kerzner, L. J. *Project Management.* Princeton, N.J.: Van Nostrand Reinhold, 1989.

Krajewski, L. J., and L. P. Ritzman. *Operations Management.* 4th ed. Reading, Mass.: Addison-Wesley, 1996.

McClain, J. O., L. J. Thomas, and J. B. Mazzola. *Operations Management.* 3rd ed. Upper Saddle River, N.J.: Prentice Hall, 1992.

Schroeder, R. J. *Operations Management.* 4th ed. New York: McGraw-Hill, 1993.

Schtub, J. F. Bard, and S. Globerson. *Project Management.* Upper Saddle River, N.J.: Prentice Hall, 1994.

Taylor, F. W. *The Principles of Scientific Management.* New York: Harper & Row, 1911.

APPENDIX **4.1**

Subprocesses and Cascading

In any given representation of a process, such as the flowchart in Figure 4.1, activities are typically treated as indivisible parts of the process. However, *any activity may be broken down further (or "exploded") into a set of subactivities;* we then refer to it as a **subprocess** of the original process. When we do this, the activity can be considered as a process in its own right, with its own set of inputs, outputs, activities, suppliers, customers, and so forth. This step can be repeated, to any level of detail desired. This begs the question as to the "right" level of detail that is to be captured by the flowchart. Obviously, the decision as to which elements of the process should be further subdivided depends on the degree of representational detail that is needed for the analysis.

In many cases, it is advantageous to depict the process at several levels of detail simultaneously. This can be achieved by using a technique called cascading. Consider, for example, the flowchart of Wonder Shed Inc. examined in Example 4.3. Assume that in a more detailed analysis of the process it was determined that Activity 4, "Form the Base", can be further divided ("exploded") into three subactivities, which are done sequentially:

Activity 4.1	Form the Front side of Base
Activity 4.2	Form the Back side of Base
Activity 4.3	Inspect the Base

In Figure 4.3 we show the original process, as well as a cascaded flowchart of the exploded activity depicted as a process in its own right.

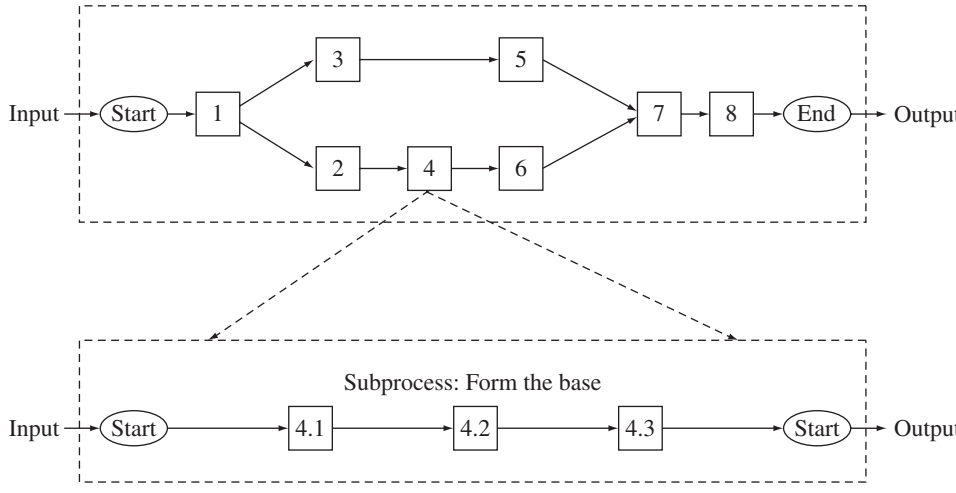


FIGURE 4.3 Cascading a Process for Wonder Shed Inc.

APPENDIX **4.2**

The Critical Path Method

In Section 4.3, we illustrated a method for finding the critical path of a process by computing the flow time of each path in the process flowchart and identifying the path with the longest flow time. However, more formal techniques have been developed for computing critical path and identifying critical activities. Here we outline one such approach, the critical path method (CPM), which is particularly useful for scheduling and controlling very large and complex projects.

Recall that all activities that lie along the critical path are labeled *critical activities* and that a delay in executing any of these activities will delay completion of the whole process. Noncritical activities, on the other hand, may be delayed (within limits) without affecting process flow time. We define the **slack time** of an activity as *the extent to which an activity could be delayed without affecting process flow time.* Thus, by definition, the slack time of a critical activity is zero. The critical path, therefore, is a path consisting of activities, all of which have a slack time of zero.

In order to compute slack time, we must calculate four time values for each activity in the processing network:

- ***Early start time (EST):*** The earliest possible time that we can begin an activity.
- ***Early finish time (EFT):*** EST plus the work content of the activity. It represents the earliest possible time we *can* finish that activity.
- ***Late finish time (LFT):*** The latest time at which an activity can end without delaying the process.
- ***Late start time (LST):*** LFT time minus the work content of the activity. It represents the latest we *must* start that activity in order not to delay its finish beyond its LFT.

Slack time may therefore be defined as follows:

$$\text{Slack time} = \text{LST} - \text{EST} = \text{LFT} - \text{EFT}$$

We now describe a systematic procedure for computing the various times. The procedure scans the process flowchart twice: forward from start to end and backward from end to start. We compute EST and EFT using **forward scheduling**—that is, *we begin at the start of the process and schedule each activity in sequence as early as possible, taking into account that it cannot start before all its predecessor activities are completed*. In contrast, LST and LFT are computed using **backward scheduling**. In that case, *we start by specifying a given target date for completing the process. We then consider each activity, in reverse order (from end to start), and compute the latest time we can complete and start this activity without violating the desired target completion time.*

COMPUTING EST AND EFT

Start to scan forward at the event "Start." Set the EST and EFT of this event to zero. Then proceed *forward,* repeating the following steps:

1. Find an activity such that the EFT of all its immediate predecessors has been computed.
2. Set the EST of the activity as the maximum of the EFTs of its immediate predecessors.
3. Set the EFT of the activity as its EST plus its work content.

The rationale behind Step 2 is that the activity can begin as soon as the last of its immediate predecessors has been completed. Once started, the EFT of the activity can be easily computed by Step 3.

We repeat Steps 1 to 3 until all the activities have been considered. The EST of the event "End" is then computed by one last iteration of Step 2. The EFT of "End" is equal to its EST and signifies the earliest time the entire unit can be completed.

COMPUTING LST AND LFT

Start the backward scan at the event "End." Set the LFT of this event to equal its EFT (computed in the forward scan). Set the LST of the "End" event to

equal its LFT. Then proceed *backward*, repeating the following steps:

1. Find an activity such that LST of all its immediate successors has been computed
2. Set the LFT of the activity as the minimum of the LSTs of its immediate successors
3. Set the LST of the activity as its LFT minus its work content

The rationale behind Step 2 is that the latest time by which an activity must be completed (without impeding the due date) is the time its successor activities must begin. In order to finish by this time, the activity must start no later than the LST computed in Step 3.

We repeat Steps 1 to 3 until all the activities have been considered. The LFT of the event "Start" is then computed (it will equal zero if the calculations were carried out correctly).

We demonstrate the forward and backward calculations for Wonder Shed (Example 4.4) in Table 4.9:

The first two columns of Table 4.9 list the activities (the number and its description) that must be performed in order to manufacture a shed. The third column gives the flow time of each activity, as in Table 4.2.

To see how the next four columns of Table 4.9 are computed, we first compute the early times, EST, and EFT using the forward scan. We start by setting the EST and EFT of the "Start" to zero. The first activity that could be analyzed is Activity 1 since all the EFTs of its predecessors (namely, the event "Start") have been computed. Thus, for Activity 1, we set its EST = 0 and EFT = 0 + 20 = 20. The next activity we can analyze is either 2 or 3 since for each of these activities the predecessor is Activity 1. Choosing Activity 2 first, we get EST = 20 (this is the EFT of activity 1). Continuing to Step 3, we get EFT = 20 + 35 = 55 for Activity 2. We continue similarly for the remaining activities.

Consider how Step 2 works for an activity with more then one predecessor, such as Activity 7. The predecessors of this activity are 5 and 6. By the time Activity 7 is considered, the EFT of these two activities was already calculated at 65 and 95, respectively. The maximum is 95, and thus the EST of Activity 7 is 95. The EFT is 95 + 15 = 110.

We continue this way for all activities, including Activity 8, whose EFT is 150. This is also the EST and EFT of the event "End."

Consider now the backward scan and the calculation of the LST and LFT. We start by setting the LST and LFT of the event "End" at 150. The first activity to schedule is Activity 8 since all its successors (namely, "End") have been scheduled. Thus, the LFT of Activity 8 is 150. We calculate its LST (Step 3) as 150 – 40 = 110. Continuing in this fashion, we ultimately reach Activity 1. Its immediate successors are 2 and 3, whose LSTs are 20 and 50, respectively. Thus, using Step 2, the LFT of Activity 1 is 20 (the minimum of 20 and 50). Step 1 yields that the LST is 20 – 20 = 0, and this is also the LST and LFT for the event "Start."

The entire calculation of the flow time, the critical path, and the slacks is summarized in Table 4.9. We see, for instance, that Activities 1, 2, 4, 6, 7, and 8 have slack times of zero and that the path connecting these activities is the critical path.

Table 4.9 Forward and Backward Calculations for Wonder Shed Inc.

	Operation	Flow Time	EST	EFT	LST	LFT	Slack Time
	"Start"	0	0	0	0	0	0
1	Separate	20	0	20	0	20	0
2	Punch the base	35	20	55	20	55	0
3	Punch the roof	25	20	45	50	75	30
4	Form the base	10	55	65	55	65	0
5	Form the roof	20	45	65	75	95	30
6	Subassemble	30	65	95	65	95	0
7	Assemble	15	95	110	95	110	0
8	Inspect	40	110	150	110	150	0
	"End"	0	150	150	150	150	0

APPENDIX **4.3**

Rework and Visits

Recall that the flow time of an activity is the time required by a typical flow unit to flow through the activity. In some situations, the process requires that activities be repeated several times before a unit is completed (e.g., due to rework). We refer to these repetitions as *visits* to an activity. For instance, if 10 percent of units require rework, we say that the number of visits to this activity is 1.1. In other situations, some units skip an operation altogether (e.g., the process requires that some, but not all, units be inspected). In that case number of visits per unit is less than one. For example, if only 10 percent of the units are inspected, the number of visits through the inspection activity is 0.1.

Define the **total flow time** of an activity to equal its *flow time multiplied by the number of visits*. Then the flow time of the process could be computed as in the previous section, by using the total flow time of the activities instead of the flow time. For instance, let us reexamine Example 4.4, the flow time for Wonder Shed Inc. Assume that malfunctioning equipment in the subassembly department has caused some quality problems in this department. Until the equipment could be fixed, 50 percent of the units need to go through this operation a second time. What is the effect of this problem on the flow time of the process? Since 50 percent of the units revisit subassembly, the number of visits is 1.5. Also, the flow time through subassembly is 30 minutes. Thus, the total flow time is thus, 1.5*30 = 45.

Since subassembly is on the critical path (Path 2), we get:

	Activity	Total Flow Time
1	Separate	20
2	Punch the base	35
4	Form the base	10
6	Subassemble the base	45
7	Assemble	15
8	Inspect	40
	Total	165 minutes

Thus, we can see that rework increases the flow time of the process from 150 minutes to 165 minutes, an increase of 10 percent.

Chapter

5

Flow Rate and Capacity Analysis

INTRODUCTION

Anticipating an explosion in the demand for cars in China—estimated to increase at over 20 percent per year in 2010 and beyond—car manufacturers are investing billions of dollars in plant, equipment, and personnel in order to expand their production capacity. Toyota, with expected sales of over 800,000 cars in 2010, is counting on its new plant in Changchun, Jilin Province, to go online in 2012 with the production capacity of 100,000 cars per year. Similarly, Nissan, a major Toyota competitor, is increasing its own capacity from less than 600,000 to 900,000 cars per year. Not to be outdone, Volkswagen is increasing it's own capacity by more than 10 percent to 850,000.

Like Toyota, Nissan, and Volkswagen, every company is striving to match its capacity to expected demand by deploying the appropriate level of resources and by maximizing the effectiveness at which these resources are utilized. This is the topic of this chapter.

In Section 5.1, we examine how flow rate and capacity can be measured. In Section 5.2, we define the effective capacity of a resource pool, and show that the effective capacity of the process depends on that of its bottleneck resources. In Section 5.3, we examine how product mix decisions impact the effective capacity of a process and its profitability. In Section 5.4, we discuss capacity waste and introduce the notions of theoretical capacity and capacity utilization. Finally, in Section 5.5, we study some key ideas to improve the capacity of a process.

5.1 FLOW RATE MEASUREMENTS

As we have defined in earlier chapters, throughput, or average flow rate, of a stable process is the average number of flow units that flow through the process per unit of time. Capacity is the maximum sustainable throughput. Throughput and capacity, which indicate a "scale" of a process, are extremely important metrics of performance: Since the flow of units through the process represents the creation of economic value, it follows that the higher the throughput, the greater the value generated by the process. Capacity is also important from the perspective of managing process flow times since insufficient process capacity may lead to congestion and excessive waiting time. For these reasons, keeping track of the flow rate of a process is one of the most fundamental tasks of management in any organization.

Throughput is expressed in terms of number of flow units per unit of time, such as customers per day, tons per shift, cars per hour, dollars per month, and patients per year. Analogous to the estimation of average flow time outlined in Chapter 4, the average flow rate (throughput) of a stable process, R, can be measured by the following three-step procedure:

1. Observe the process over a given, extended period of time.
2. Measure the number of flow units that are processed by the process over the selected period of time.
3. Compute the average number of flow units per unit of time.

As mentioned earlier, the capacity is the maximum sustainable flow rate. It can be measured by observing the system in periods of heavy congestion in which the flow rate is limited by (and therefore equal to) capacity.

The lean operations literature, which began with the Toyota Production System (Chapter 10), often describes throughput in terms of **takt time**. Derived from the German word for rhythm or beat, takt time is the *reciprocal of throughput.* The concept is particularly useful in the context of synchronized assembly lines, where it represents the average activity time at each workstation (takt times for the assembly of mass-produced cars are on the order of 1 minute). Takt time is sometimes also called cycle time, but some authors use cycle time as a synonym for flow time. To avoid confusion, we do not use the term cycle time in this book.

5.2 RESOURCES AND EFFECTIVE CAPACITY

In general, the capacity of a system depends on the level of resources deployed by the system and on the effectiveness at which these resources are utilized. The capacity of any given process is typically quite difficult to analyze, mainly due to the subtle and complicated ways in which the various resources can interact. In this section we provide a simple and useful approximation, called the effective capacity.

5.2.1 Resources and Resource Pools

As we discussed in Chapter 1, activities are performed by capital and labor resources. Each activity may require one or more resources, and each resource may be allocated to one or more activities. For example, in the process of making bread, raw materials—flour, salt, butter, water, and so forth—are transformed into loaves of bread. The entire process requires performing such activities as mixing the ingredients, kneading the dough, forming the loaves, placing them in the oven(s), and baking them. In turn, these activities use such resources as mixers, bakers, and ovens. A given resource—for instance, a baker—may be used by several activities, such as mixing, kneading, and forming dough. Similarly, a given activity, such as loading the oven, may require multiple resources, such as a baker and an oven.

A **resource pool** is *a collection of interchangeable resources that can perform an identical set of activities. Each unit in a resource pool* is called a **resource unit**. For instance, in the case of bread making, three flexible bakers, each of whom can perform any baking activity, would be viewed collectively as a single resource pool containing three resource units. On the other hand, if each of the three bakers specialized in a separate activity (mixing, kneading, and forming dough, respectively), they would be regarded as three separate resource pools, each consisting of a single resource unit.

Combining separate resource pools into a single, more flexible, pool able to perform several activities is called **resource pooling**. It is a powerful operational concept that can significantly affect not only process flow rate and process capacity but also flow time (as we will see in Chapter 8).

5.2.2 Effective Capacity

The **unit load of a resource unit** is the *average amount of time required by the resource unit to process one flow unit, given the way the resource is utilized by the process*. The **effective capacity of a resource unit** is the *inverse of the unit load.* It represents the maximum sustainable flow rate through the resource unit, if it were to be observed in isolation. The **effective capacity of a resource pool** is the *sum of the effective capacities of all the resource units in that pool.* As an illustration, consider an insurance agent (a resource unit) whose job is to file residential insurance claims (flow units). Assume, for example, that on average, the agent spends 12 minutes per claim (unit load). Then, the effective capacity of the resource unit is 1/12 claims per minute (60/12 = 5 claims per hour). If two agents were available to process claims, then the effective capacity of the resource pool would be $2 \times 5 = 10$ claims per hour.

Formally, if we denote the unit load (at resource pool i) by T_i and the number of resource units by c_i

$$\text{Effective capacity (of resource pool } i\text{)} = c_i/T_i \qquad \textbf{(Equation 5.1)}$$

In practice, the agent is likely to handle various types of claims which require different amounts of time. The unit load in this case represents the average amount, over all types of claims. This issue is taken up in detail in Section 5.3. Also, if the various resource units are not identical in terms of their effective capacities, then the effective capacity of the resource pool will be the sum of the effective capacities of each resource unit in the pool. For the rest of this chapter, however, we will assume that all units in a resource pool are identical.

Effective capacities of different resource pools may vary. Since all resource pools are required to process each flow unit, no process can produce output any faster than its **bottleneck**—*the "slowest" resource pool of the process.* Thus, we define the **effective capacity of a process** as *the effective capacity of the bottleneck.*

The effective capacity of a process is a very useful concept for managing capacity, as Equation 5.1 provides a simple and practical way for connecting process capacity to overall resource levels, given the effectiveness at which resources are employed.

EXAMPLE 5.1

Health maintenance organizations (HMOs) provide their customers with all-inclusive medical service for a fixed monthly fee. To secure services, they contract with physicians and hospitals that provide their services on a fee-per-service basis. When members of an HMO receive medical service, the providing physician or hospital submits a claim to the HMO for reimbursement. NewLife Finance is a service provider to HMOs. For a small fee, it performs the entire claims processing operation on behalf of the HMO.

Table 5.1 Effective Capacity for Physician Claims, NewLife Finance

Resource Pool (i)	Unit Load (minutes per claim) (T_i)	Effective Capacity of a Resource Unit (claims per minute) ($1/T_i$)	Number of Units in the Resource Pool (c_i)	Effective Capacity of a Resource Pool (claims per minute) (c_i/T_i)
Mailroom clerk	1.00	1/1 = 1.00	1	1.00
Data-entry clerk	5.00	1/5 = 0.20	8	1.60
Claims processor	8.00	1/8 = 0.125	12	1.50
Claims supervisor	2.50	1/2.5 = 0.40	5	2.00

Processing a physician claim consists of the following operations:

1. Claims billed by physicians arrive by mail and are opened and date-stamped by the mailroom clerk. They are then placed into a data-entry bin.
2. Data-entry clerks enter date-stamped applications—first in, first out—into NewLife's claims-processing system. Data-entry clerks must check claims for proper formatting and completeness of data fields before they input claims into the system. If a claim is not legible, fully completed, or properly formatted, it must be sent back to the physician for resubmission. Once entered, claims are stored in a processing inventory called "suspended claims."
3. Claims are assigned to a claim processor for initial processing.
4. Processed claims are transferred by the system to a claim supervisor for inspection and possible alterations.
5. Claims are returned to their original claim processors who complete the transaction and issue instructions to accounts payable for settlement.

The process involves five steps, performed by four resource pools, namely mailroom clerks, data entry clerks, claim processors, and claim supervisors (Steps 3 and 5 are performed by the same resource pool, the claims processors). Table 5.1 lists, for each of the four resource pools, the unit load, its inverse, that is, the effective capacity of the resource unit, and the number of resource units. Based on this information, the last column of the table computes the effective capacity of each of the resource pools.

As can be seen, the bottleneck of the process is the pool of mail room clerks; and the effective capacity of the entire system equals 1.00 claim per minute, or 60 claims per hour.

How many professionals are required to achieve capacity of, say, 80 claims per hour? Using the effective capacity as our guideline, we note that an additional "third" (20/60) of a mailroom clerk is necessary. If part-time solutions are not available, one additional clerk must be hired. That will increase the capacity of the mailroom to 120 claims per hour, which is more than sufficient. Note, however, that the capacity of the *process* will not increase to 120: The bottleneck in this case shifts to the pool of claims processors, which has just enough capacity to handle 90 claims per hour.

5.2.3 Capacity Utilization

The throughput of a process, R, is the average number of flow units processed over a given period of time. Throughput of a process may not equal capacity because of external constraints such as low outflow rate (due to low demand rate, meaning an external bottleneck in the output market) or low inflow rate (due to low supply rate, meaning an external bottleneck in the input market).

For the i th resource pool, the **capacity utilization** of the resource pool, denoted by u_i, is defined by the relation:

u_i = Throughput/effective capacity of the i th resource pool **(Equation 5.2)**

Capacity utilization indicates the extent to which resources—which represent invested capital—are utilized to generate outputs (flow units and ultimately profits). It is defined for each resource pool independently. The capacity utilization of the process is defined by the bottleneck resource pool. We illustrate the concept in Example 5.2:

EXAMPLE 5.2

Assume that the average number of claims processed by NewLife Finance during a given month was measured to be 400 per day. The effective capacity of the various resources is as indicated in the second column of Table 5.2. The third column of the table lists the capacity utilization of the various resources:

Table 5.2 Capacity Utilization for NewLife Finance

Resource Pool (p)	Effective Capacity of a Resource Pool (claims per 8-hour day)	Capacity Utilization (u_i)
Mailroom clerk	$1.00 \times 480 = 480$	400/480 = 83%
Data-entry clerk	$1.6 \times 480 = 768$	400/768 = 52%
Claims processor	$1.5 \times 480 = 720$	400/720 = 56%
Claims supervisor	$2.0 \times 480 = 960$	480/960 = 42%

Notice that, by definition, the bottleneck resource is the most highly utilized resource. If the throughput were to increase, that resource will be the first to hit full utilization. The capacity utilization of the entire process is 83 percent, given by the bottleneck.

5.2.4 Extensions: Other Factors Affecting Effective Capacity

The calculations of effective capacity discussed in the previous section ignore several important factors. First, they assume that resources handle units sequentially, or one unit at a time, rather than in load batches (imagine loaves of bread baked in an oven). Second, they assume that all resources are available for the same amount of time (imagine a factory with some units running a second shift). Finally, we have ignored the effects of setups or switching between products (imagine an operating room that needs to be reset between different types of surgery). Often these assumptions do not hold. We discuss how Equation 5.1 can be adjusted to accommodate these factors in Appendix 5.1.

5.3 EFFECT OF PRODUCT MIX ON EFFECTIVE CAPACITY AND PROFITABILITY OF A PROCESS

Firms often produce several products simultaneously. Since various products utilize resources at different rates, the effective capacity depends on the products produced and their proportions in the mix. This observation has an important business implication. In most organizations, sales/marketing departments make product mix decisions. Since such decisions affect the process capacity (a major driver of profitability), input from the operations group, which is responsible for production, is required. In Example 5.3,

Table 5.3 Unit Loads for Various Products, NewLife Finance

Resource Pool	Unit Load (Physician) (minutes per claim)	Unit Load (Hospital) (minutes per claim)
Mailroom clerk	1.00	1.50
Data-entry clerk	5.00	6.00
Claims processor	8.00	8.00
Claims supervisor	2.50	4.00

Table 5.4 Effective Capacity for Hospital Claims, NewLife Finance

Resource Pool (i)	Unit Load (minutes per claim) (T_i)	Effective Capacity of a Resource Unit (claims per minute) ($1/T_i$)	Number of Units in the Resource Pool (c_i)	Effective Capacity of a Resource Pool (claims per minute) (c_i/T_i)
Mailroom clerk	1.50	1/1.50 = 0.66	1	0.66
Data-entry clerk	6.00	1/6.00 = 0.17	8	1.33
Claims processor	8.00	1/8.00 = 0.125	12	1.50
Claims supervisor	4.00	1/4.00 = 0.25	5	1.25

we demonstrate the dependence of capacity on the product produced. In Sections 5.3.1 and 5.3.2, we examine the issue of product mix.

EXAMPLE 5.3

Assume that, in addition to processing physician claims, NewLife also handles claims submitted directly by hospitals. The process used to handle hospital claims is the same process used for physician claims. However, the unit loads required for the various operations are different. Table 5.3 contrasts the unit loads required for the two types of products.

In Table 5.4, we recompute the effective capacity of the process, using the unit loads that pertain to hospital claims (second column). The calculations of the other columns follow the logic of Example 5.1.

Thus, the capacity of the process is only 0.66 claims per minute (40 per hour) for hospital claims, as opposed to 60 claims per hour for physician claims.

5.3.1 Effective Capacity for Product Mix

For a process that produces several types of products simultaneously, we can represent the overall flow of the various products by constructing an (artificial) flow unit which represents the entire mix of the various products. We can calculate the unit load of the mix by averaging the unit loads of the individual products, using the weights of the mix, as illustrated in Example 5.4:

EXAMPLE 5.4

Currently, NewLife handles a product mix of 60% physician claims and 40% hospital claims. What is the effective capacity of the process?

Table 5.5 Unit Loads for Various Products, NewLife Finance

Resource Pool	Unit Load (Physician) (minutes per claim)	Unit Load (Hospital) (minutes per claim)	Unit Load (60%–40% mix) (minutes per claim)
Mailroom clerk	1.00	1.50	1.20
Data-entry clerk	5.00	6.00	5.40
Claims processor	8.00	8.00	8.00
Claims supervisor	2.50	4.00	3.10

Table 5.5 lists again the unit loads of the two types of claims. The fourth column, listing the unit loads of the 60%–40% mix, is computed by taking the weighted average of the previous two columns. For example, the unit load for the mailroom clerk equals $60\% \times 1.0 + 40\% \times 1.5 = 1.20$.

We can compute the effective capacity of the product mix in the same way as that of an individual product, using the averaged unit load instead of the individual values. This is illustrated in Table 5.6:

Table 5.6 Effective Capacity for 60%–40% Product Mix, NewLife Finance

Resource Pool (i)	Unit Load (minutes per claim) (T_i)	Effective Capacity of a Resource Unit (claims per minute) ($1/T_i$)	Number of Units in the Resource Pool (c_i)	Effective Capacity of a Resource Pool (claims per minute) (c_i/T_i)
Mailroom clerk	1.20	1/1.20 = 0.83	1	0.83
Data-entry clerk	5.40	1/5.40 = 0.185	8	1.48
Claims processor	8.00	1/8.00 = 0.125	12	1.50
Claims supervisor	3.10	1/3.10 = 0.32	5	1.61

As can be seen, the effective capacity in this case is 0.83 claims per minute, or 50 claims per hour. As expected, this falls between the effective capacities of the individual products (1 and 0.66 claims per minute respectively). However, the effective capacity of the mix is *not* equal to the 60%–40% weighted average of the respective effective capacities. Rather, it is the unit load, the inverse of the effective capacity, which is equal to the 60%–40% weighted average of the respected total unit loads.

In the case of NewLife, the pool of mailroom clerks is the bottleneck resource pool for every product mix. However, in general, a change in the product mix can affect not only the effective capacity but also the bottleneck.

5.3.2 Optimizing Profitability

Which of the two claims processed by NewLife Finance—physicians or hospitals—is more profitable? To answer this question we need to supplement the data on capacities for the two products with financial information concerning revenues and variable costs.

The **unit contribution margin** of each flow unit is *its revenue less all of its variable costs.* For instance, if the revenues per unit for physician and hospital claims are $5.50 and $6.75 per unit respectively, and that the variable costs are $0.5 and $0.75, then the unit contribution margins for the two products are $5.5-0.5 = \$5$ and $6.75-0.75 = \$6$ per unit, respectively.

On first sight it may seem that the product with the highest unit contribution margin is the most profitable, and thus one may conclude that hospital claims are more profitable than physician claims ($6 per unit is more than $5 per unit). However, as we see below, assessing the profitability of products solely on the basis of contribution margin per unit ignores the essential role that capacity plays in determining profitability. This is demonstrated in Example 5.5:

EXAMPLE 5.5

Table 5.7 summarizes the information about the two products:

Table 5.7 Effective Capacity and Contribution Margins, NewLife Finance

	Physician Claims	Hospital Claims
Effective Capacity (units per hour)	60	40
Contribution margin ($ per unit)	5.00	6.00

Consider first the case of processing only physician claims. Since the capacity is 60 per hour, and the margin per unit is $5, we can generate $60 \times 5 = \$300$ per hour. On the other hand, if we process only hospital claims, we can generate $40 \times 6 = \$240$ per hour. Thus, even though the contribution margin per unit for hospital claims is larger, it is the less profitable product. Clearly in this case, the fact that we can make a higher profit on each unit of hospital claims is more than offset by the fact that we can process fewer units per hour.

As the example demonstrates, the relevant criterion in determining the profitability of products is not the contribution *per unit* but the contribution *per unit of time*. In essence, this metric corresponds to viewing capacity and throughput in terms of financial flows rather than in terms of flow of physical units ($300 of contribution margin per hour rather than 60 claims per hour). The concept combines the relevant financial information, namely contribution margin per unit, with the operational concepts of capacity and throughput.

In practice, product mix problems are likely to be larger, more complicated, and involve other considerations, such as demands for the various products and other marketing constraints. In Appendix 5.2 we briefly present a general methodology, called Linear Programming, for handling such issues.

5.4 CAPACITY WASTE AND THEORETICAL CAPACITY

In most cases, the routine operation of a process involves a considerable amount of capacity waste due to factors such as resource breakdown, maintenance, quality rejects, rework and repetitions, setups between different products or batches, non-value-adding activities, and so forth. The unit load, which is used to determine effective capacity, is an aggregation, given the way the resources are currently being utilized, of the "productive" as well as the "wasted" time. However, if capacity waste is large, we may want to turn our attention to waste elimination; and thus, it is useful to "segregate" the wasted capacity. We discuss this issue in this section.

5.4.1 Theoretical Capacity

The **theoretical unit load of a resource unit** is *the minimal amount of time required to process a flow unit, if all waste were eliminated*. The **theoretical capacity of the resource unit** is *the reciprocal of the theoretical unit load.* It represents the maximum sustainable

flow rate through the resource unit, if it were to be utilized without any waste. The **effective capacity of a resource pool**, and of the entire process, is defined as in Section 5.2.2. Similar to the concept of theoretical flow time examined in Chapter 4, the **theoretical capacity of a process** provides *a highly idealized, and seldom attainable, notion of capacity*. Its usefulness derives from the fact that it provides an estimate of the waste in the system and forms the basis for any action plan for waste elimination.

In some cases, the theoretical unit load could be observed or estimated directly. If this is not possible, one could sometimes estimate instead the amount of capacity waste in the system and compute the theoretical unit load (and theoretical capacity) using a waste factor. This is demonstrated in Example 5.6:

EXAMPLE 5.6

Consider the operating room (a resource unit) of a hospital which specializes in cataract surgery. On average, the hospital manages to perform a surgery every 30 minutes. This is the unit load. Thus, the effective capacity is 2 cases per hour.

Suppose, it is estimated that on average 33 percent of the operating room time is wasted (cleaning, restocking, changeover of nursing staff, fixing of malfunctioning equipment and so forth). Thus, the theoretical unit load can be estimated as $30 \times (1–33\%) = 20$ minutes, yielding a theoretical capacity of 3 cases per hour.

Formally, let CWF indicate the capacity waste factor, expressed as a percentage of the theoretical capacity. Then,

$$\text{Theoretical Capacity} = \text{Effective Capacity}/(1 - \text{CWF}) \qquad \textbf{(Equation 5.3)}$$

Note that the capacity waste factor may vary among the various resource pools.

5.4.2 Theoretical Capacity Utilization

In a similar way to the definition of capacity utilization introduced earlier, we can define the **theoretical capacity utilization of a resource pool** as the

Throughput/theoretical capacity of the i *th resource pool.*

The theoretical capacity utilization may be of interest since it includes the effects of internal inefficiencies, in addition to capacity loss due to external factors. The theoretical capacity utilization of the process is defined by the bottleneck resource pool.

5.5 LEVERS FOR MANAGING THROUGHPUT

Now let us examine some of the levers available for managing the throughput of a given process. As mentioned earlier, such levers have a powerful impact on profitability. Since a large fraction of the costs of operating a process are fixed, *small changes in throughput could be translated into large changes in profits*. We call this magnification the **throughput profit multiplier**. Formally,

$$\text{Throughput profit multiplier} = \%\text{ change in profit}/\%\text{ change in throughput} \qquad \textbf{(Equation 5.4)}$$

This effect is illustrated in Example 5.7:

EXAMPLE 5.7

Consider a process with the following economics of operations. The fixed costs of owning and operating the resources amount to $180,000 per month. The revenue is $22 per unit and the variable costs amount to $2 per unit. Thus, the contribution margin

is \$20 per unit. In July 2010, the process throughput was 10,000 units. The profit for July was then:

$$\$20 \text{ per unit} \times 10{,}000 \text{ units} - \$180{,}000 \text{ fixed costs} = \$20{,}000 \text{ profit}$$

A process improvement team was able to increase output in August by 1% to 10,100 units, without any increase in the fixed cost. The profit for August was then:

$$\$20 \text{ per unit} \times 10{,}100 \text{ units} - \$180{,}000 \text{ fixed costs} = \$22{,}000 \text{ profit}$$

Thus, a 1% increase in throughput has resulted in a 10% increase in profits—a throughput profit multiplier of 10!

The throughput profit multiplier could be computed directly, as in Example 5.7. It could also be obtained by the formula:

$$\text{Throughput profit multiplier} = \text{contribution margin per unit} / \text{profit per unit}. \quad \textbf{(Equation 5.5)}$$

For example, for the data of Exercise 5.7, margin per unit is \$20 and profit per unit is \$20,000 per month/10,000 units per month = \$2 per unit. Using Equation 5.5, we get a throughput profit multiplier yielding a factor of $20/2 = 10$.

If the throughput profit multiplier is large, the financial impact of increasing throughput is significant. But what are the most effective ways to increase throughput? We discuss this question in the following sections.

5.5.1 Throughput Improvement Mapping

Before any throughput improvement project is initiated, it is useful to get *a view of the big picture and identify the most likely source of additional throughput*. We call this activity **throughput improvement mapping**. As a starting point, consider the relationships among the various concepts we have examined so far.

$$\text{Throughput} \leq \text{Capacity} \leq \text{Theoretical Capacity} \quad \textbf{(Equation 5.6)}$$

(We may substitute the term "Capacity" in Equation 5.6 with "Effective Capacity" which is the approximation for capacity used in this chapter.)

When throughput is significantly less than capacity, we say that the "bottleneck is external"—the process is limited by factors that lie outside its bounds, such as the demand for its outputs or the supply of its inputs. In that case, the only way to increase output is to increase the capacity of this external bottleneck. In the case of demand bottleneck, this could be accomplished, for instance, by lowering prices, increasing quality levels, increasing sales efforts or increasing the advertising budget. If the bottleneck involves supply, we may need to identify additional suppliers, modify some of the supply chain processes and so on.

If the throughput is about equal to capacity, we say that the "bottleneck is internal." In this case the only way to increase throughput is by increasing capacity. This can be done in two ways. First, we can increase the financial capacity of the process, by modifying the product mix. This topic was covered in Section 5.3.2. Alternatively, we can increase the physical capacity of the process. In that case, there is additional useful information we can obtain from the relationships in Equation 5.6. Specifically, if capacity is about equal to theoretical capacity, then existing resources are very efficiently utilized, and extra capacity will require increasing the level of resources. This is examined in section 5.5.2. On the other hand if capacity is significantly lower than the theoretical capacity, then the existing resources are not utilized effectively, and the key to extra throughput is the elimination of waste. This is covered in section 5.5.3.

5.5.2 Increasing Resource Levels

If process capacity is about equal to the theoretical capacity, the key to increasing capacity is to increase the theoretical capacity of the process. This requires increasing the resource levels of each bottleneck resource pool. This could be achieved by taking some of the following actions:

1. Increase the number of resource units
2. Increase the size of resource units
3. Increase the time of operation
4. Subcontract or outsource
5. Speed up the rate at which activities are performed

In each case, the improvement should be directed towards the bottleneck resource. We now elaborate on these levers in more detail:

Increase the Number of Resource Units: Adding more units of the resource to the bottleneck resource pool will increase its theoretical capacity. Naturally, the cost of these resources must be compared to the value of the extra throughput generated. Therefore, this alternative is particularly appealing when the bottleneck resources are relatively cheap, readily available, and easy to install.

Increase the Size of Resource Units: Because resources can often process multiple units simultaneously—a phenomenon referred to as load batching—one simple way to increase resource capacity is to increase the load batch of the resource. For example, if we have an oven that can bake ten loaves at a time, we could increase its capacity by replacing it with a bigger oven that can accommodate fifteen loaves at a time. The effect of the load batch on capacity is discussed further in Appendix 5.1.

Increase the Time of Operations: Extending the scheduled availability, that is the time period during which the bottleneck resource operates, will increase the theoretical capacity. In both manufacturing and service operations, increasing the hours of operation and employee overtime are common methods to increase process output. The effect of scheduled availability on capacity is also further discussed in Appendix 5.1.

Subcontract or Outsource Bottleneck Activities: Instead of buying more units of a bottleneck resource, one could subcontract or outsource similar capacity. This is a very common way to increase capacity but may involve higher operating and coordinating costs.

Speed Up the Rate at which Activities Are Performed: Decreasing the time it takes the bottleneck to perform an activity will result in an increase in capacity. This approach is often of limited effectiveness and may involve investments in faster resources or incentives to workers.

5.5.3 Reducing Resource Capacity Waste

Reducing capacity waste on the bottleneck is one of the most effective ways to increase capacity. Some of the most common ways to achieve this are listed here:

1. Eliminate non-value-adding activities
2. Avoid defects, rework, and repetitions
3. Reduce time availability loss
4. Reduce setup loss
5. Move some of the work to nonbottleneck resources
6. Reduce interference waste

Note the similarity of levers 1 and 2 to levers mentioned in Chapter 4. The only difference is in our choice of activities to improve: Whereas for flow time we focus on activities along the critical path, for improving flow rate we focus on activities performed by bottleneck resources. In the following section, we elaborate on the remaining levers 3 to 6:

Reduce Time Availability Loss: Breakdowns, work stoppage and other interruptions reduce the time available for processing units. They can be reduced by improved maintenance policies, by scheduling preventive maintenance outside periods of availability, and by effective problem-solving measures that reduce the frequency and duration of breakdowns.

Reduce Setup Waste: Time used for setups is wasted as far as throughput is concerned. This is discussed further in Appendix 5.1. Setup waste can be reduced by decreasing the frequency of changeovers, working proactively to reduce the time required for each setup, and managing the product mix. When we decrease the frequency of changeovers, however, we need to produce a larger number of units of a product (batch size) before changing over to produce a different product. As noted in Chapter 4, this increased batch size, however, may lead to higher inventories and longer flow times (we will discuss this topic further in Chapter 6).

Move Some of the Work to Nonbottleneck Resources: Removing some work off a bottleneck resource frees some time to process additional units. This may require greater flexibility on the part of nonbottleneck resources as well as financial investments in tooling, cross-training, and so forth. However, since nonbottleneck resources have, by definition, excess capacity, this may be an excellent way to gain capacity.

Reduce Interference Waste: Interference waste occurs due to interactions and lack of synchronization among the various resources. For example:

- **Starvation**: Resources are available but cannot process units because some of the necessary inputs are unavailable.
- **Blocking**: Resources are prevented from producing more flow units because there is no place to store the already processed flow units or additional processing has not been authorized.

In both cases, the problem does not lie in the resource unit itself but is caused by issues generated elsewhere. Starving can be minimized by allowing for an adequate buffer before (upstream from) the bottleneck resource. Similarly, blocking can be reduced by placing a large buffer following (downstream from) the bottleneck. Proper buffer size selection is discussed in Chapters 7 and 8. Additionally, starvation and blocking can be reduced by synchronizing the flow throughout the process. Synchronization is discussed in detail in Chapter 10.

5.5.4 Shifting Bottlenecks and the Improvement Spiral

As we have seen earlier, the only way to increase the throughput of a process is to identify its bottleneck (external or internal) and to increase its capacity. Once the capacity of the bottleneck is increased above a certain level, however, the bottleneck shifts—the original bottleneck is no longer the bottleneck—and it is replaced by some other resource whose capacity is now the lowest. Once this happens, it is futile to increase the capacity of the "old" bottleneck any further since it is no longer a bottleneck. The only way to increase the capacity further is to shift attention to the new bottleneck and increase its capacity. Therefore, in selecting the level of financial investment in resources, we should look closely at the process capacity that we are trying to improve. As we relax bottlenecks by adding more resource units, new bottlenecks may appear, and the total process

capacity may increase, but only at a decreasing rate. In particular, the bottleneck can shift from internal to external several times as we spiral through these steps. For an extensive treatment of the throughput improvement spiral, see Goldratt (1990).

Summary

In Chapter 3, we introduced three measures of process performance, namely flow time, flow rate, and inventory, and discussed the relations among them (Little's law). In Chapter 4, we studied the first of these measures, namely, the flow time. In this chapter, we examined the flow rate of a process.

The throughput, or average flow rate of a stable process, is the average number of flow units that flow through the process per unit of time. Capacity is the maximum sustainable throughput of a process. The throughput is a key operational measure of performance since the process creates value by its stream of flow units, and thus the higher the throughput, the higher the value created in a given period. Capacity is also important from the perspective of flow times since insufficient process capacity may lead to excessive waiting time.

The capacity of a process depends on its resource levels and on the efficiency at which these resources are utilized. The effective capacity of a process serves as a simple approximation for the capacity of a process and allows us to study the connection between capacity and resource levels. A key concept in computing the effective capacity is the bottleneck. Capacity utilization is the ratio of throughput to effective capacity.

The theoretical capacity represents the capacity of the process, if resources could be utilized without any waste. It represents an aspiration level for the throughput of the process that can never be achieved in practice. The theoretical capacity utilization is a measure of how close a given process is to this ideal.

The throughput of a process is limited by either an internal (within the process) or an external bottleneck. When the bottleneck is external, throughput can be improved by increasing sales effort or improving the supply of inputs. When the bottleneck is internal, throughput can be increased by increasing process capacity. This can be done by increasing financial capacity, by giving priority to the more profitable products, or by increasing the physical capacity.

Physical capacity can be increased by increasing resource levels or by reducing capacity waste.

Key Equations and Symbols

(Equation 5.1) Effective capacity of a resource pool $i = (c_i/T_i)$

(Equation 5.2) Capacity utilization u_i = Throughput/effective capacity (of the i th resource pool)

(Equation 5.3) Theoretical Capacity = Effective Capacity/(1 − CWF)

(Equation 5.4) Throughput profit multiplier = % change in profit/% change in throughput

(Equation 5.5) Throughput profit multiplier = contribution margin per unit/profit per unit

(Equation 5.6) Throughput ⩽ Capacity ⩽ Theoretical capacity

where

T_i = Unit load at resource pool i

c_i = Number of resource units in resource pool i

CWF = Capacity Waste Factor

u_i = Capacity utilization of resource pool i

Key Terms

- Blocking
- Bottleneck
- Changeover
- Capacity utilization
- Effective capacity of a process
- Effective capacity of a resource pool
- Effective capacity of a resource unit
- Load batching
- Resource pool
- Resource pooling
- Resource unit
- Scheduled availability
- Setup
- Setup batch
- Starvation
- Takt time
- Theoretical capacity of a process
- Theoretical capacity of the resource unit
- Theoretical capacity utilization of a resource pool
- Theoretical unit load of a resource unit
- Throughput improvement mapping
- Throughput profit multiplier
- Total unit load
- Unit contribution margin
- Unit load of a resource unit

Discussion Questions

5.1 While visiting your favorite restaurant, identify the major resource pools and the bottleneck and estimate the overall capacity utilization.

5.2 Explain the concepts of unit load and theoretical unit load for an airline such as Southwest Airlines.

5.3 The theoretical capacity of a process is the reciprocal of the theoretical flow time of the process. Do you agree? Explain.

5.4 List examples of service organizations that rely mainly on setup reductions to improve their capacity and throughput.

5.5 List examples of organizations that rely on judicious product mix decisions in order to maximize their throughput and revenues.

5.6 Comment on the statement, "To maximize profitability, it is always better to give priority to produce products with the highest unit contribution margins."

5.7 Comment on the statement, "Doubling the number of units of a bottleneck resource will double the process capacity."

5.8 Comment on the statement: "Maximizing utilization of each resource pool is an exercise in futility."

Exercises

5.1 A law firm specializes in the issuance of insurance policies covering large commercial real estate projects. The projects fall into two categories: shopping centers, and medical complexes. The typical work involved in each transaction is quite predictable and repetitive. The time requirements (unit loads) for preparing a standard contract of each type are given in Table 5.8. Also listed are the number of professionals of each type and the number of available hours per professional per day (the rest of time is taken by other office activities):

Table 5.8 Data for law Firm

	Unit Load Shopping (hours per contract)	Unit Load Medical (hours per contract)	No. Of Professionals	Hours Available (hours per professional per day)
Paralegal	4	6	4	6
Tax lawyer	1	3	3	8
Senior partner	1	1	2	4

For the month of November, 2010, the firm has generated the 150 orders, 75 of each type. Assume one month equals 20 days.

- **a.** What is the effective capacity of the process (contracts per day)?
- **b.** Can the company process all 150 cases in November?
- **c.** If the firm wishes to process all the 150 cases available in November, how many professionals of each type are needed?

***5.2** Reconsider the law firm of Exercise 5.1. Assume the prevailing revenues per shopping and medical projects are $4000 and $5000 per project, respectively, and that out of pocket expenses associated with each project are negligible. The (fixed) cost of operating the office is $500,000 per month.

- **a.** What type of project is the most profitable?
- **b.** At the current product mix (50%-50%), how much contribution margin is generated ($ per day)?
- **c.** At the current product mix, what is the profit at capacity?
- **d.** At the current product mix, what is the value of hiring an extra Paralegal?

***5.3** Three hairstylists, François, Bernard, and Mimi, run Fast Service Hair Salon for busy professionals in the Gold Coast area of downtown Chicago (see Figure 5.1). They stay open from 6:45 a.m. to 9:00 p.m. in order to accommodate as many people's work schedules as possible. They perform only shampooing and hairstyling activities. On average, it takes 10 minutes to shampoo, 15 minutes to style the hair, and 5 minutes to bill the customer. When a customer arrives, he or she first checks in with the receptionist (Bernard's younger sister LuLu). This takes only 3 minutes. One of the three stylists then takes charge of the customer and performs all three activities—shampooing, styling, and billing—consecutively.

- **a.** What is the number of customers that can be serviced per hour in this hair salon?
- **b.** A customer of Fast Service Hair Salon, an operations specialist, has suggested that the billing operation be transferred to LuLu. What would be the impact on the theoretical capacity?

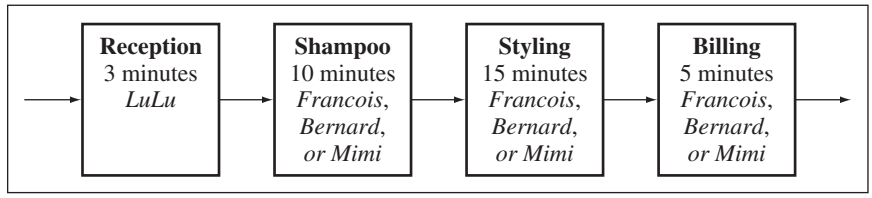


FIGURE 5.1 Current Process at Fast Service Hair Salon

5.4 A company makes two products A and B, using a single resource pool. The resource is available for 900 minutes per day. The contribution margins for A and B are $20 and $35 per unit respectively. The unit loads are 10 and 20 minutes per unit.
- **a.** Which product is more profitable?
- **b.** The company wishes to produce a mix of 60% As and 40% Bs. What is the effective capacity (units per day)?
- **c.** At the indicated product mix, what is the financial capacity (profit per day)?

5.5 An insurance company processes two types of claims: Life and Property. The capacity of processing life claims is 500 per month. The capacity of processing property claims is 1000 per month.
- **a.** Assuming a common bottleneck, what is the capacity of processing a mix of 50%-50% of the two types?

***5.6** A company's average costs and revenues for a typical month are $15 million and $18 respectively. It is estimated that 33% of the costs are variable, and the rest is fixed.
- **a.** What is the throughput profit multiplier?

5.7 Reexamine Exercise 5.1. Assume that the capacity waste factors of the paralegals, tax lawyers, and senior partners are 20%, 30%, 35%, respectively.
- **a.** What is the theoretical capacity of the process?

Selected Bibliography

Eppen, G. D., F. J. Gould, C. P. Schmidt, J. H. Moore, and L. R. Weatherford. *Introductory Management Science.* 5th ed. Upper Saddle River, N.J.: Prentice Hall, 1998.

Goldratt, E. M. *Theory of Constraints.* Croton-on-Hudson, N.Y.: North River Press, 1990.

Goldratt, E. M., and J. Cox. *The Goal.* 2nd rev. ed. Barrington, Mass.: North River Press, 1992.

Winston, W. L. *Operations Research: Applications and Algorithms.* 2nd ed. Boston: PWS-Kent Publishing, 1991.

APPENDIX **5.1**

Other Factors Affecting Effective Capacity: Load Batches, Scheduled Availability, and Setups

In this appendix we review how Equation 5.1 should be modified to include the effects of resource size, time of operations, and setups. We handle the first two:

Load Batches Often, *resources can process several flow units simultaneously, a phenomenon referred to as* **load batching**. For example, consider an oven that can bake 10 loaves simultaneously. We say in that case that its load batch is 10. The computations of effective capacity used in Section 5.2 continue to apply if we just define a flow unit as one batch. Naturally, the higher the load batch, the higher the capacity.

Scheduled Availability Typically, each resource unit is scheduled for operation only a portion of the total time (e.g., eight hours per day, five days per week). *The amount of time that a resource is scheduled for operation is called the* **scheduled availability** *of the resource.* Scheduled availability of various resources in a process may differ. For example, in a manufacturing setting, it is not uncommon that some areas within a plant operate only one shift per day (8 hours) while others operate two (16 hours). Moreover, the choice of one day as the time period of measurement is based on the assumption that availability patterns repeat on a daily basis. However, more complicated patterns are possible. Some resource pools, for example, may be available only on Mondays and Thursdays, with the pattern repeating every week. In that case, we should measure scheduled availability in number of hours per week.

It is easy to modify the expression for effective capacity to account for the effects of the load batch and scheduled availability. Let

LB_i be the load batch of resource pool i

SA_i be the scheduled availability of resource pool i

Then,

$$\text{Effective capacity (of resource pool } i) = (c_i/T_i) \times LB_i \times SA_i$$

For example, consider a resource pool containing two ovens, each of which bakes 10 loaves of bread simultaneously (load batch). The baking time (unit load) is 15 minutes. Finally, assume that the oven is scheduled for operations 7.5 hours (450 minutes) per day. The effective capacity of the pool is given as $(2/15) \times 10 \times 450 = 600$ loaves per day.

Consider now the effect of setups. In a process that involves multiple products, it may be necessary to set up the process each time the product is changed. The *cleaning, resetting, or retooling of equipment in order for it to process a different product is called* **setup** or **changeover**. For instance, the painting robots in a paint shop require draining the pipelines and cleaning of the painting heads each time the color is changed. Similarly, when researchers in a research-and-development organization switch among several research projects, they are likely to waste time with every such switch.

We denote the average time required to set up a resource (at resource pool i) for a particular product by S_i (minutes per setup). Assume that once the setup is completed, we run a continuous batch of Q_i units of the same product before we change over and set

Table 5.9 Total Unit Loads and Effective Capacity

Q (units)	Total unit Load (minute per unit)	Capacity (units per hour)
5	10+ 60/5 = 22	60/22= 2.7
10	10+ 60/10 = 16	60/16= 3.75
20	10+ 60/20 = 13	60/13= 4.6
60	10+ 60/60= 11	60/11= 5.4
120	10+ 60/120 = 10.5	60/10.5= 5.7

up the resource for the next product. We refer to Q_i as the **setup batch** or the lot size—*the number of units processed consecutively after a setup.*

Since we utilize one setup for Q_i units, the average time for setup per unit is S_i/Q_i. This time should be added to the *unit load* of the product to yield the **total unit load**

$$\text{Total unit load} = T_i + S_i/Q_i$$

To compute effective capacity, we replace unit load with total unit load

$$\text{Effective capacity (of resource pool } i) = c_i/(T_i + S_i/Q_i)$$

For example, assume that the unit load of a given product, T_i, is 10 minutes per unit and that the setup time, S_i, is 60 minutes per batch. In Table 5.9 we summarize, for various sizes of setup batch, Q_i, the total unit load and the effective capacity.

What is the "right" lot size or the size of the setup batch? On the one hand, the higher we set the lot size, the lower the total unit load will be and thus the higher the capacity. On the other hand, the higher we set the lot size, the higher the inventory will be and consequently (using Little's law) the higher the flow time. We explore this relationship and also the determination of the optimal batch size in detail in Chapter 6.

Note that the setup batch discussed here is an entirely different concept than the load batch introduced earlier in this appendix. Whereas the load batch relates to the ability of the resource to handle units *simultaneously,* as in the case of baking bread, the setup batch indicates the number of units processed *sequentially,* between subsequent setups. The load batch is often constrained by the technological capabilities of the resource, such as the size of the oven. In contrast, the setup batch is determined managerially—it simply represents the number of flow units of a given type that are processed before we switch over to a different type. In the paint shop example discussed earlier, the setup batch represents the number of units of a particular color that are painted before we switch the process to another color.

APPENDIX **5.2**

Optimizing Product Mix with Linear Programming

Determining the optimal product mix can be stated as a problem of allocating resources to products in order to maximize profits. Consider, for example, the case of NewLife Finance, Inc. Assume, that in addition to Physician and Hospital claims considered in Section 5.3, the company could also process a third type of claims, called Government claims. Each claim requires work at four resource pools: mailroom clerks, data entry clerks, claim processors, and claim supervisors.

The data needed for the analysis are, for each type of claim, the contribution margin per unit, and the unit load at each resource pool. Also needed is the number of resource units at each resource pool. Finally, we need information about the maximum demand of each type of claim. The data is summarized in Tables 5.10 and 5.11:

If we denote the number of the three types of claims, that is, Physician, Hospital, and Government by P, H, and G, respectively (claims per hour), then we can formulate the problem as follows:

1. Maximize $5P + 6H + 4.5G$

Such that:

2. $1P + 1.5H + 2G \leq 60$
3. $5P + 6H + 5G \leq 480$
4. $8P + 8H + 10G \leq 720$
5. $2.5P + 4H + 4.5G \leq 300$
6. $P \leq 30$
7. $H \leq 50$
8. $G \leq 40$
9. P, H, and G are nonnegative

The expression in (1) represents the financial throughput, which we wish to maximize. The Inequalities (2 to 5) ensure that we do not spend, in any department, more time than we have. Note that

Table 5.10 Unit Loads for NewLife Finance

		Unit load (minutes per claim)		
Resource	**No. of units**	**Physician**	**Hospital**	**Government**
Mail Room Clerks	1	1.0	1.5	2.0
Data Entry Clerks	8	5.0	6.0	5.0
Claims Processors	12	8.0	8.0	9.0
Claims Supervisors	5	2.5	4.0	4.5

Table 5.11 Unit Contribution Margins and Demand for NewLife Finance

	Physician	**Hospital**	**Government**
Contribution Margin ($ per unit)	5	6	4.5
Demand (units per hour)	30	50	40

the right hand side in each case equals the number of resource units multiplied by 60 minutes. The inequalities (6 to 8) represent the demand constraints. Finally, (9) is required since there is no meaning to negative throughput.

We can solve the problem using the Solver routine of Microsoft Excel, or using any linear programming package (see Winston, 1991). The optimal solution is found to be

$$P = 30$$
$$H = 20$$
$$G = 0$$

Linear programming will also indicate that the optimal profit is 270 (dollars per hour) and that the bottleneck is the mail room clerk.

In the simple case of NewLife Finance, it could be easily established that the bottleneck is the mail room clerk, for any product mix. Also, the most profitable product is P, followed by H and finally G. Therefore, NewLife should process as many physician claims as possible: 30 per hour in this case. Since each claim requires 1 minute of the bottleneck resource, this will consume 30 minutes per hour. The next profitable product is H. NewLife should process as many of these as possible, given the time left for the mailroom clerks. Since the unit load is 1.5 minute per claim, this amount to 20 hospital claims. There is no time life in the mail room to handle G, the least profitable product.

In practice, optimization problems are much larger and more complicated. The formal optimization technique of linear programming can be used for easily solving these optimal product mix problems. Such techniques are detailed in many operations research and management science textbooks such as those listed at the end of this chapter.

Chapter 6

Inventory Analysis

INTRODUCTION

The health care industry faces enormous pressures to improve quality of care and access to care in an environment of declining reimbursements and increasing costs. Surveys reveal that a typical hospital spends between 25 to 30 percent of its budget on medical, surgical, and pharmaceutical supplies. The procurement process has traditionally relied on manual ordering, reconciliation, and payment processes resulting in an effective cost of cutting a purchase order that is often higher than the product being purchased. Furthermore, surveys reveal that different areas within a hospital appear to order their own supplies leading to bloated and costly inventory. There appears to be a tremendous opportunity to streamline the supply chain and improve procurement practices to reduce inventory.

In the past few years, several hospitals have begun to look at their materials management systems to improve efficiency. Phoebe Putney Health System in Georgia is expecting to save $3 million over five years. The Memorial Sloan Kettering Institute is using the Internet to control procurement costs. Centura Health, a nine-hospital integrated delivery network in Denver, is using e-commerce to trim and improve on its $100 million annually spent on supplies. Several of these initiatives require substantial investments in information

technology and warehousing. Before making such investments, a hospital's materials management staff needs to understand the key drivers of inventory.

Health care is not the only industry plagued by inventory. While inventory is ubiquitous across all industries, the ability of companies to manage them varies dramatically within and across industry sectors. Developing better inventory management capability can significantly affect the bottom line. Consider, for example, the retail book industry. Borders Group, Barnes & Noble, and Amazon are the three largest booksellers in the United States. In 2009, the Borders Group with annual sales of $2791 million and gross margins of 21.5 percent, carried about 145 days of inventory in its network. In contrast, Barnes & Noble carried about 121 days of inventory in 2009. If the Borders Group could improve its inventory management capability to match the 121 days of inventory of Barnes & Noble, it would reduce its working capital requirement by approximately $147.5 million.

From a macroeconomic perspective, inventory-related costs accounted for approximately 2.5 percent of the gross domestic product of the United States in 2009. According to the U.S. Department of Commerce, which tracks monthly sales and inventory for the U.S. economy, in 2009 the average monthly inventory in the U.S. economy was about $1.37 trillion on annual sales of about $12 trillion. Of this, the inventory at the manufacturer, wholesaler, and retailer levels was $522 billion, $403 billion, and $411 billion, respectively, which shows that there is enormous opportunity to make significant impact by better inventory management.

What are the various reasons for carrying inventory? What is the impact of procurement transaction costs on inventory? What is the value of aggregating purchases across multiple entities? What are the right metrics? How does better inventory management affect the bottom line?

In this chapter and the next, we provide a framework to answer these questions. We begin with an analysis of different types of inventories, reasons for carrying them, and the associated costs. Then we analyze the key trade-offs in managing inventory under an economies-of-scale effect. In Chapter 7, we discuss the protective role of inventories against unforeseen changes in demand or supply.

We begin in Section 6.1 with a broad classification of inventory depending on its location in the process and introduce the concept of theoretical inventory. In Sections 6.2 and 6.3, we identify the reasons for carrying inventories and the various costs of holding inventories. Section 6.4 derives the inventory dynamics under batch purchasing and processing. Section 6.5 examines the optimal **inventory level** that balances costs and benefits. Section 6.6 studies the effect of lead times on ordering decisions. Section 6.7 describes inventory implications for a firm that follows a periodic inventory policy. Finally, in Section 6.8, we conclude the chapter by summarizing some key levers for managing various types of inventory. Appendix 6.1 shows derivation of the economic order quantity formula and Appendix 6.2 discusses the topic of price discounts and their effect on inventory.

6.1 INVENTORY CLASSIFICATION

In addition to flow time and flow rate (throughput), which we studied in Chapters 4 and 5, inventory is the third basic measure of process performance. As we did with the other two measures, we first identify the boundaries of the process under study. Then we define inventory as the number of flow units present within those boundaries. Because average inventory is related by Little's law to both average flow time and average flow rate, controlling inventory allows us to indirectly control flow rate, flow time, or both. Inventory also directly affects cost—another important measure of process performance. Because it affects several dimensions of process performance, inventory is a key lever in managing business process flows.

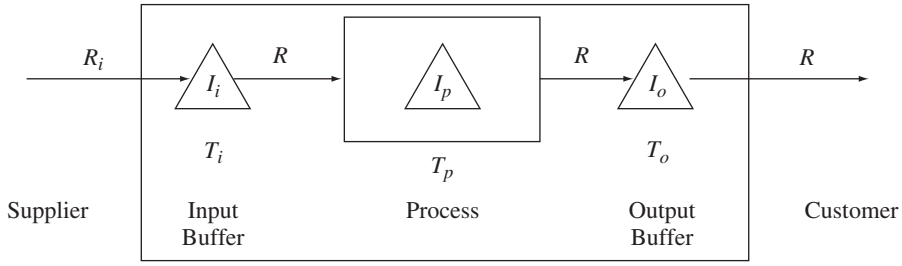


FIGURE 6.1 Process Flows and Inventories

Inventory includes all flow units within the process boundaries. Depending on the inventory's location or stage in the process, we can classify units of inventory as belonging to one of three categories:

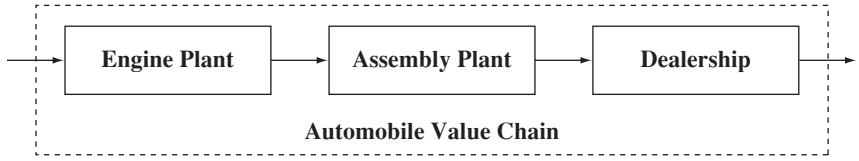
1. *Flow units that are waiting to begin processing constitute* **input inventory**.
2. *Flow units that are being processed constitute* **in-process inventory**.
3. *Processed flow units that have not yet exited process boundaries accumulate in* **output inventory**.

Figure 6.1 shows the process flow and the three stages of inventory accumulation.

In-process inventory can further be classified as work-in-process or in-transit inventory. **Work-in-process inventory** *are the flow units being processed in a manufacturing or service operation.* **In-transit inventory** or **pipeline inventory** *refers to the flow units being transported*.

In a manufacturing process, input inventory consists of raw materials or components, in-process inventory includes all work being processed, and output inventory contains finished goods. The classification of inventory, however, will also depend on where the process boundaries are drawn. Output inventory for one process can be input inventory for the other, as illustrated in Figure 6.2.

In a service process, a flow unit is typically a customer. Input inventory here refers to customers waiting for service, and in-process inventory refers to customers being served. If served customers leave the process immediately, there is no output inventory. We will postpone the analysis of the special problems of managing inventory of customers in service processes to Chapter 8.



Process	Input Inventory	In-Process Inventory	Output Inventory
Engine plant	Castings	Unfinished engines	Finished engines
Assembly plant	Finished engines, chassis, etc.	Unfinished automobile	Automobile
Dealership (sales)	Automobile	Automobile	—

FIGURE 6.2 Inventory Classification for an Automobile Value Chain

We begin by establishing the following notation:

$$\text{Average input inventory} = I_i$$
$$\text{Average in-process inventory} = I_p$$
$$\text{Average output inventory} = I_o$$

Thus, average total inventory, I, within process boundaries can be expressed as

$$I = I_i + I_p + I_o$$

A flow unit moving through the process will then spend some time in each of three classes of inventory. Average values of these waiting times are denoted as follows:

$$\text{Average time spent in input inventory} = T_i$$
$$\text{Average time spent in in-process inventory} = T_p$$
$$\text{Average time spent in output inventory} = T_o$$

Total average flow time, therefore, can be expressed as

$$T = T_i + T_p + T_o$$

If we denote average process flow rate in equilibrium as R, then flow units enter and leave each stage at this rate R. As we have seen in Chapter 3, Little's law applies to the aggregate values, giving the relationship $I = R \times T$. Little's law can also be applied to each of the stages to establish the relationship between the flow rate in equilibrium and the corresponding average inventory and average time at that stage.

Theoretical Inventory Although Little's law determines average inventory, an imbalance between inflows and outflows that develops over time will cause actual inventory to fluctuate around this average. In Chapter 3, we discussed in detail the inventory dynamics and also introduced the concept of the inventory buildup diagram. Briefly, inventory accumulates at any stage in a process whenever inflow into that stage exceeds the outflow from that stage. Similarly, inventory depletes at any stage whenever the outflow from a stage exceeds the inflow into that stage.

However, even in an ideal situation with perfectly balanced flows—one in which inflow, processing, and outflow rates are all equal at every point in time—we still encounter in-process inventory. Recall from Chapter 4 the concept of theoretical flow time, which represents the minimal flow time in a process. Even if no flow unit ever waits in a buffer, it remains within the process boundaries as work in process until it exits the process. Therefore, if a process needs to produce some output, there will always be some inventory within its boundaries. To remain consistent with the concepts of theoretical flow time and theoretical capacity that we discussed earlier, we introduce the concept of theoretical inventory and denote it by I_{th}. Much like theoretical flow time, **theoretical inventory** is *the minimum amount of inventory necessary to maintain a process throughput* of R and can be expressed as

$$\text{Theoretical inventory} = \text{Throughput} \times \text{Theoretical flow time}$$

$$I_{th} = R \times T_{th} \qquad \textbf{(Equation 6.1)}$$

Theoretical inventory is the average inventory for a given throughput if no flow unit ever had to wait in any buffer. It represents the minimal amount of flow units undergoing activities (without waiting) to sustain a given flow rate. Like theoretical flow time, theoretical inventory gives us an optimal target to aim for.

In reality, of course, flow units may wait in buffers before being processed at any stage, leading to a flow time longer than the theoretical flow time. Consequently, in-process inventory will often be larger than the theoretical inventory.

Decoupling Processes Input and output inventories form buffers that decouple the process from its environment, thereby permitting relatively independent operations. Input inventory permits the process manager to manage processing rates independently of material inflow (supply) rates; output inventory permits managing the processing rate independently of product outflow (demand) rate.

Input and output inventories may be viewed and analyzed in the same way—each has its supplier and customer, and each serves as a buffer between the two. If inflow (supply) into the buffer exceeds outflow (demand) from the buffer, the excess is added to the buffer; if outflow exceeds inflow, the buffer shrinks. If the buffer is emptied, the next stage in the process is "starved" of work. Such starvation, which we mentioned briefly in Chapter 5, typically deteriorates process performance. For example, starvation in output inventory results in stockouts and customer dissatisfaction, and starvation in in-process or input inventory results in lost production. Starvation occurs because of unpredictable events that affect the inflow to or outflow from the buffer. Management of buffer inventories to prevent starvation will be discussed in detail in Chapter 7. In general, several factors may affect the buildup and build-down of inventory in buffers, giving rise to various reasons for holding inventories, which we discuss next.

6.2 INVENTORY BENEFITS

Why do firms carry inventory? As we already observed, a minimum level of in-process inventory, called theoretical inventory, is necessary to maintain a given process throughput. Reducing inventories to less than the theoretical inventory will result in a loss of throughput. In transportation and logistics, flow units are transported from one location to another. The units that are being transported (that are en route) at a given point in time constitute in-transit or pipeline inventory. Pipeline inventory is necessary to allow the functioning of a business process whose activities are distributed over many locations. In practice, however, firms plan and maintain far in excess of the theoretical and pipeline inventory. Therefore, the question is: Why do firms intentionally plan for such excesses? We now survey four possible answers to this question.

6.2.1 Economies of Scale

We say that a process exhibits **economies of scale** when *the average unit cost of output decreases with volume.* Economies of scale may arise from either external or internal causes in areas such as procurement, production, or transportation. One reason firms intentionally plan for such excess inventory is to take advantage of economies of scale, thereby making it attractive to procure, produce, or transport in quantities more than immediately needed. If, for example, an external supplier offers price discounts, the buyer may find it economical to procure in quantities larger than those needed for immediate processing. Internally, perhaps the buyer finds it more economical to procure or process in large quantities because of a fixed cost that is incurred each time the activity is undertaken. For example, procuring input often involves a **fixed order cost**—*the administrative cost of processing the order, transporting the material, and receiving and inspecting the delivery.* Each of these costs may add a significant fraction to total cost that is independent of order size. For example, if a truck is dispatched each time an order must be picked up, a large fraction of the cost of the trip will be independent of the quantity ordered (up to the size limit of the truck). In producing output, the process of starting production runs may involve a **fixed setup cost**—*the time and materials required to set up a process* (e.g., clean equipment and change tools). An ice cream maker, for instance, must clean vessels before changing from chocolate to vanilla. The time taken for changeovers is unavailable for production, thus decreasing throughput. Hence, the

manager may decide to produce large quantities of an ice cream flavor before switching over to produce another flavor. Chapter 5 highlighted the impact of fixed setup times on flow rate and flow time. In this chapter, we will study its impact on inventory.

We often refer to *the order or production in response to the economies of scale effect as a* **batch**. Sometimes, more specific names are associated with it, depending on the type of activity being performed, such as a **production batch** (or *lot*) *in the case of production,* a **transfer batch** *for transportation or movement,* and **procurement batch** (or *order*) *for purchasing quantities*. The number of units constituting a batch is called a **batch size**, also referred to as lot size. Faced with an economies-of-scale effect, a manager finds it more economical to procure or produce infrequently in large batches, thereby spreading the fixed cost over more units. Such practice of intermittent procurement or production leads to periodic buildup and build-down of inventory, as we will see in the next section. *The average inventory arising due to a batch activity is referred to as* **cycle inventory**.

6.2.2 Production and Capacity Smoothing

A related reason for planning inflows in excess of outflows is production and capacity smoothing. When demand fluctuates seasonally, it may be more economical *to maintain a constant processing rate and build inventories in periods of low demand and deplete them when demand is high.* This is often referred to as a **level-production strategy**. Examples of leveling include building toy inventories throughout the year for sales during the Christmas season or producing lawn mowers year-round for spring and summer sale. Demand fluctuations are then absorbed by inventories rather than by intermittent and expensive adjustments in processing capacity. Likewise, there could be seasonality in supply. For example, agricultural products are the key inputs to the food processing industry. The flow of agricultural inputs into a processing center will exhibit seasonality that is dependent on the timing of harvests. To maintain a level production, then, input inventories are allowed to accumulate and are then gradually depleted over time. *Inventories that serve as buffers to absorb seasonal fluctuations of supply and demand are called* **seasonal inventories**. Although the costs of holding of inventory increase, a level production strategy minimizes the cost of capacity changes.

An alternate strategy to deal with demand fluctuations is called **chase demand strategy**, *whereby a firm produces quantities exactly to match demand.* By matching demand and production, the firm, of course, carries no inventory. Unfortunately, it also transfers all demand fluctuations to its processing system. In particular, matching demand patterns would require constantly altering process capacity or its utilization with associated costs.

Which is the better strategy—level production or chase demand? The answer depends on the relative magnitudes of the fixed costs of altering capacity and the variable costs of holding inventory. It is better to level production if capacity changes are expensive and to chase demand if inventories are expensive to carry. Not surprisingly, the true optimum lies somewhere between these two extremes, employing a combination of capacity adjustments and inventory buffers. The problem of finding this optimal combination is called **aggregate production planning**. A detailed discussion of this topic is beyond the scope of this book, and we refer the reader to Chopra and Meindl (2009) or Nahmias (2008).

6.2.3 Stockout Protection

The third reason for holding inventories is to protect against stockouts due to unexpected supply disruptions or surges in demand. Any number of events—supplier strikes, fire, transportation delays, and foul weather—may reduce input availability. Potential consequences to the buyer include process starvation, downtime, and temporary

reduction in throughput. Many producers, therefore, maintain inventories of inputs to insulate the process and continue operation despite supply shortages.

Likewise, because customer-demand forecasts are usually inaccurate, planning process output to meet only forecasted demand may result in stockouts, delayed deliveries, lost sales, and customer dissatisfaction. Thus, many producers maintain cushions of excess inventory to absorb excess demand and ensure product availability and customer service despite forecast errors. *Inventory maintained to insulate the process from unexpected supply disruptions or surges in demand is called* **safety inventory** or **safety stock**. In Chapter 7, we will discuss the degree of stockout protection provided by a given level of safety inventory and the level of safety inventory needed to provide a desired level of protection.

6.2.4 Price Speculation

The fourth reason for holding inventories is to profit from probable changes in the market prices of inputs or outputs. In addition to protecting against sudden price changes due to such crises as wars or oil shortages, speculative inventories of commodities (such as corn, wheat, gold, and silver) and financial assets (such as stocks, bonds, and currencies) can be held as investments. As prices fluctuate over time, investors can manage their inflows (purchases) and outflows (sales) to optimize the financial value of their inventories. In the semiconductor industry, for instance, a rapid price decline of chips over time gives computer manufacturers a very strong incentive to delay purchasing chips as inputs; rather, they can wait to enjoy the latest, and often lowest, purchase price. The process manager then holds some **speculative inventory**. Although speculative inventories are important in finance and economics, we will not study them in any detail in this book, as our focus is more on processing operations.

To illustrate the remaining concepts in this chapter, we will use the example of procurement decisions in a hospital network described in Example 6.1.

EXAMPLE 6.1

Centura Health[1] is a nine-hospital integrated delivery network based in the Denver area in the United States. Currently each hospital orders its own supplies and manages the inventory. A common item used is a sterile Intravenous (IV) Starter Kit. Weekly demand for the IV Starter Kit is 600 units. The unit cost of an IV Starter Kit is $3. Centura has estimated that the physical holding cost (operating and storage costs) of one unit of medical supply is about 5 percent per year. In addition, the hospital network's annual cost of capital is 25 percent. Each hospital incurs a fixed order cost of $130 whenever it places an order, regardless of the order size. The supplier takes one week to deliver the order. Currently, each hospital places an order of 6,000 units of the IV Starter Kit whenever it orders. Centura has recently been concerned about the level of inventories held in each of the hospitals and is exploring strategies to reduce them. The director of materials management is considering the following options:

1. Increase the frequency of ordering by reducing the current order size
2. Centralize the order process across all nine hospitals and perhaps serving all the hospitals from a single warehouse

[1]All numbers are fictitious and are used only to illustrate the concepts.

6.3 INVENTORY COSTS

Carrying inventory is expensive both in operational and in financial terms. Assume that a firm is carrying a large inventory of work-in-process and outputs. If market demand shifts to new products, the firm is left with two choices. One is to empty the process by scrapping all current work-in-process and liquidating the obsolete outputs inventory at marked-down prices and then quickly introducing the new product. This option results in a significant loss on the old inventory. The other choice is to finish processing all in-process inventory and sell all output before introducing the new product. This option causes delay in the launch of the new product, which creates a reduced responsiveness to the market.

Large inventories also delay execution of design changes because current inventory must be processed and sold first. Moreover, the buildup of inventories between successive processing stages has other operational consequences. For example, it obstructs workers' view of the total process. It also decouples the operation of consecutive stages of processing such that each stage works independently of the other. Such decoupling, however, may discourage teamwork and coordination across a process. We will discuss these operational inefficiencies from holding inventories further in Chapter 10.

Inventory Holding Cost Carrying inventory entails a financial cost called **inventory holding cost**, *which has two components—the physical holding cost and the opportunity cost of capital tied up in inventory:*

1. **Physical holding cost** refers to *the cost of storing inventory.* It includes all operating costs (insurance, security, warehouse rental, lighting, and heating/cooling of the storage) plus all the costs that may be entailed before inventory can be sold (spoilage, obsolescence, pilferage, or necessary rework). The former costs are largely fixed. Physical holding cost per unit of time (typically a year), however, is usually expressed as a fraction h of the variable cost C of acquiring (or producing) one flow unit of inventory. Thus, the physical holding cost of carrying a unit of inventory for one time unit is hC.
2. **Opportunity cost** *of holding inventory refers to the forgone return on the funds invested in inventory which could have been invested in alternate projects.* Indeed, inventory shows up as an asset on the balance sheet because it is an economic resource that is expected to be of future value. The firm could realize this value by liquidating it and investing the proceeds elsewhere. Or, even more to the point, the sooner inventory sells, the sooner it creates accounts receivable—and the sooner accounts receivable generates cash. The opportunity cost of holding one flow unit is usually expressed as rC, where r is the firm's rate of return (measured as annual percentage return on capital) and C is the variable cost of acquiring (or producing) one flow unit of inventory (measured as cost/flow unit).

Together, the physical and opportunity costs of inventory give the total unit inventory holding cost per unit time, denoted by H, which is expressed as follows:

$$\begin{aligned}\text{Total unit inventory holding cost} &= \text{Unit physical holding cost} \\ &\quad + \text{Unit opportunity cost of capital}\end{aligned}$$

$$H = (h + r)C \qquad \textbf{(Equation 6.2)}$$

For example, if the unit time period is a year, H represents the total cost of keeping one unit in inventory for one year. Example 6.2 illustrates the computation of H for Centura Health introduced in Example 6.1.

EXAMPLE 6.2

Consider Centura Health introduced in Example 6.1. Recall the unit cost of the IV Starter Kit, $C = \$3$. Furthermore, the annual physical holding of 5% implies

$$hC = \$(0.05)(3) = \$0.15/\text{year}$$

Centura's annual cost of capital is $r = 25\%$. Thus, a dollar of inventory carries an opportunity cost of \$0.25 per year in terms of possible alternate uses of the funds. Since \$3 is tied up in each unit of the IV Starter Kit, the opportunity cost of keeping one unit in inventory for a year is

$$rC = \$(0.25)(3) = \$0.75/\text{year}$$

Hence, the total annual holding cost of a unit of IV Starter Kit is

$$H = (h + r)C = \$0.15 + \$0.75 = \$0.90$$

While H represents the unit inventory holding cost per unit time, the total inventory holding cost per unit time will be $H \times I$, where the average inventory level is I. Therefore, to decrease the holding cost of inventory, we have two levers:

1. Decrease unit inventory holding cost H (typically by getting concessions on h or C).
2. Decrease average inventory level I.

6.4 INVENTORY DYNAMICS OF BATCH PURCHASING

In the rest of this chapter, we focus on the effect of economies of scale, discussed in Section 6.2.1, leading to what we call cycle inventory. We start with developing the inventory profile resulting from batch purchasing. We will rely on the basic methodology outlined in Chapter 3 to study the dynamics of flow rate and inventory. Subsequently, we will present a framework for decision making under economies of scale. While we focus our discussion on purchasing, the concepts apply equally well to other batch activities.

We illustrate these concepts as a means of answering two important managerial questions that arise at a hospital in Centura Health, introduced in Example 6.2, during the process of purchasing IV Starter Kits (inputs). Recall that Centura buys a batch of IV Starter Kits at a time even though the consumption of these kits has been quite steady for the past several months. The hospital materials manager at a Centura hospital must decide (1) *how much* to purchase at a time, and (2) *when* to purchase a new batch of IV Starter Kits. Both decisions will affect the hospital's balance sheet as they impact costs.

To answer these questions, we analyze the inventory dynamics of the procurement process for the IV Starter Kit at Centura hospital. We must consider the following procedures that are valid in this purchasing scenario:

1. Inputs are procured in multiple units at a time—a system called batch purchasing.
2. Processing (or consumption) takes place continuously at a constant rate.

As an answer to the first question, how much to purchase at a time, we will assume that the manager buys the IV Starter Kit in batches of Q units at a time and analyze the financial consequences of such a decision. In the next section, we will determine the optimal quantity to purchase. The entry point into the process is the point at which a batch of Q units is delivered to a Centura hospital and added to its inputs inventory. Assume that the hospital has a steady consumption rate of R and the initial input inventory, just before the first IV Starter Kit delivery arrives at time $t = 0$, is zero.

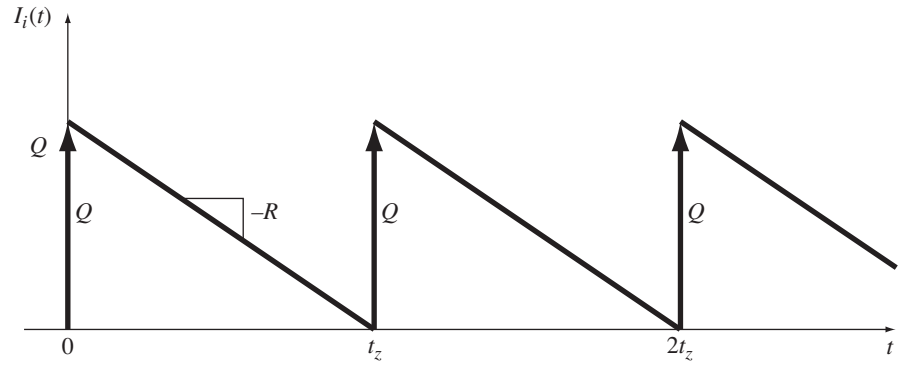


FIGURE 6.3 Inventory Profile with Batch Size *Q*

As inputs inventory will vary over time, we will denote its level at any time t with $I_i(t)$. Thus, just after the first batch is received at time 0, we have inputs inventory, $I_i(0) = Q$. After the first delivery, inflow rate remains at zero until the next shipment is received. Outflow rate due to consumption, meanwhile, remains constant at R. After the first delivery, therefore, the process inventory buffer is depleted at a rate R, so that change in flow rate, denoted by ΔR, is negative R, or $\Delta R = -R$. Consequently, only input inventory is depleted at rate R, whereby it will reach zero after, say, time t_z, so that $I_i(t_z) = 0$. Thus, we have

$$\text{Input inventory at time } t_z = \text{Input inventory at time } 0 - \text{Total demand during time } t_z$$

$$I_i(t_z) = Q - R \times t_z = 0$$

so that

$$t_z = Q/R$$

Simply stated, it takes Q/R time units to deplete input stock of size Q at rate R. If the process manager always orders in batches of size Q, the same cycle repeats itself every t_z time units. Over time, the resulting inventory buildup diagram displays the sawtooth pattern shown in Figure 6.3. It answers our second question, *when* to order a new batch of IV Starter Kits: Centura hospital should order another batch of IV Starter Kits to arrive whenever the total inventory drops to zero (and thus the input buffer is empty). As a result, Centura should place orders so that a batch arrives every t_z time units.

Under batch purchasing and a constant depletion rate, therefore, the input inventory profile is triangular with height Q. In a typical order cycle, average input inventory is one half the batch (or order) size or

$$I_i = Q/2$$

In terms of flow time, the first flow unit purchased in each batch is consumed immediately, while the last unit spends all the $t_z = Q/R$ time units in input inventory buffer storage before its use can begin. Thus, an average flow unit spends $t_z/2 = (Q/\mathrm{R})/2 = Q/2R$ time units in input inventory storage. Alternately, we can apply Little's law to determine the average flow time spent in the input buffer as

$$T_i = (Q/2)/R = Q/2R$$

To summarize, the cyclical pattern of inventory buildup and build-down with a batch size of Q gives us an average input inventory of $Q/2$. While we focused our discussion on inputs inventory, batching could lead to similar patterns in in-process or outputs inventory buffers. Thus, the average inventory of $Q/2$ is driven primarily by batching

and is not particular to input, in-process, or output buffers. We label the average inventory arising due to batch size of Q as cycle inventory and denote it by

$$I_{cycle} = Q/2 \qquad \textbf{(Equation 6.3)}$$

Example 6.3 illustrates the situation at Centura Health.

EXAMPLE 6.3

Recall that one of the hospitals of Centura Health processes a demand of 600 units of the IV Starter Kit each week and places an order of 6,000 kits at a time. The hospital must then be ordering once every 10 weeks. Accordingly, average IV Starter Kit inventory will be $I_{cycle} = 6{,}000/2 = 3{,}000$ units, and a typical IV Starter Kit spends an average of five weeks in storage. Thus, the Centura hospital carries an average cycle inventory of 3,000 IV Starter Kits.

Whenever there are economies of scale, it is cost effective to order (or produce) in batches. Although our focus has been on a purchasing process with flow units of products, it equally applies to other processes and types of flow units, as illustrated in the following examples:

- Joe Smith goes to withdraw cash from an automated teller machine (ATM). Since he has to drive two miles to get to the closest ATM for his bank, he prefers to withdraw sufficient cash to meet his entire week's cash expenses. This practice of periodic withdrawal leaves a cycle inventory of cash in Joe's wallet.
- Office Assistants trains people for secretarial and office administrative tasks for placement with its clients. A team of experts conducts the program. The compensation paid to this team is usually independent of the size of the class. Hence, it is considered a fixed cost. Office Assistants decides to restrict the class size to 45 people who, upon training, are absorbed in the workforce in about three months. The average inventory of the pool of trained but unemployed people constitutes cycle inventory.
- Big Blue runs a campus shuttle service at the University of Michigan. There are fixed costs of running a bus along a specific route. Therefore, the university finds it economical to traverse the route with a limited frequency. The number of people who arrive between two consecutive trips becomes the batch size. The average number of people waiting at a bus stop to board a bus is cycle inventory.
- The City of Pittsburgh collects trash from its residents' homes once every week. Meanwhile, trash accumulates in the garbage cans every day until it gets picked up on Monday. The average inventory of trash in the household constitutes that household's cycle inventory.

Like the purchasing process, the first two examples illustrate a situation where there is an instantaneous buildup of inventory followed by a gradual depletion of it. The next two examples, unlike the purchasing process, involve a situation with a gradual buildup of inventory followed by an instantaneous build-down. The resulting inventory profile will be different from that of Figure 6.3. Both situations, nevertheless, lead to cycle inventory because of the periodic nature of the respective activities.

6.5 ECONOMIES OF SCALE AND OPTIMAL CYCLE INVENTORY

Process managers would like to determine inventory levels that optimally balance costs and benefits of carrying various inventories. In the remainder of this chapter, we will show how to determine the optimal level of cycle inventory that balances the costs of

holding inventory with the benefits of economies of scale. In doing so, we will distinguish two causes of scale economies:

1. Economies arising from a fixed-cost component of costs in either procurement (e.g., order cost), production (e.g., setup cost), or transportation
2. Economies arising from price discounts offered by suppliers

In the main chapter, we will concentrate on the former and postpone analysis of price discounts to Appendix 6.2.

Fixed Cost of Procurement: Economic Order Quantity As mentioned earlier, our analysis applies equally to input and output buffers. Indeed, in batch procurement or purchasing, we analyze input buffers, while in batch processing we analyze in-process and output buffers. Suppose outflow from the buffer occurs at a constant rate of R flow units per unit of time (e.g., per year). Assume that the process manager can control inflow into the buffer. Each time the inflow is initiated by the procurement of material, a fixed cost of ordering, S, is incurred regardless of the quantity procured. It is therefore more economical to procure inputs infrequently in batches, even though outflow requirements remain steady over time. Let Q be the size of each order (batch) procured at any given time. The annual order frequency, which represents the number of times we need to order to satisfy an annual outflow rate of R, is

$$\begin{aligned}\text{Annual order frequency} &= \text{Annual outflow rate/Order size}\\ &= R/Q\end{aligned}$$

Since each order incurs a fixed cost S, the total annual fixed order cost is

$$\text{Fixed cost per order} \times \text{Annual order frequency}$$

or

$$S \times R/Q$$

Observe that this annual order cost decreases as the order size Q increases because the more we order at a given time, the fewer orders we need to place over the course of a year.

Conversely, recall from Section 6.4 that average cycle inventory is $I_{cycle} = Q/2$ units. Consequently, total annual inventory holding cost is expressed as

$$\text{Unit holding cost per year} \times \text{Average inventory}$$

or

$$H \times I_{cycle} = H \times (Q/2)$$

Note that this cost increases when order size Q increases.

Finally, we must also consider total annual cost of materials procured, which is given by

$$\text{Unit cost} \times \text{Outflow rate}$$

or

$$C \times R$$

Observe that the annual cost of materials is independent of the choice of order size Q. We assume the unit variable cost to be constant; that is, we get no price discounts for purchasing in large quantities. We discuss price discounts in Appendix 6.2.

Thus, the total annual cost, denoted by TC, is given by the sum of the total annual fixed order cost, total annual inventory cost, and total annual cost of materials as follows:

$$TC = S \times \frac{R}{Q} + H \times \frac{Q}{2} + C \times R \qquad \textbf{(Equation 6.4)}$$

Example 6.4 illustrates the cost structure of the current operating policy at a Centura Health hospital.

EXAMPLE 6.4

Recall that a Centura Health hospital incurs a cost of \$130 regardless of the quantity purchased each time it places an order for and receives IV Starter Kits. Hence, fixed cost per order $S = \$130$. From Example 6.1, we also know that unit cost $C = \$3$, and the weekly outflow rate of 600 translates to an annual outflow of $R = 31{,}200$ per year, assuming 52 weeks per year.

In Example 6.2, we computed Centura's inventory holding cost as $H = \$0.90$ per unit per year. Recall that a Centura hospital currently procures 6,000 units in each order, so we have $Q = 6{,}000$. The components of the total annual cost can be computed as

$$\begin{aligned}\text{Total annual fixed order cost} &= S \times R/Q \\ &= 130 \times 31{,}200/6{,}000 = \$676 \\ \text{Total annual holding cost} &= H \times (Q/2) \\ &= 0.90 \times 3{,}000 = \$2{,}700 \\ \text{Total annual cost of materials} &= C \times R \\ &= 3 \times 31{,}200 = \$93{,}600\end{aligned}$$

The total annual cost can thus be computed as

$$\begin{aligned}TC &= S \times \frac{R}{Q} + H \times \frac{Q}{2} + C \times R \\ &= 676 + 2{,}700 + 93{,}600 \\ &= \$96{,}976\end{aligned}$$

In fact, once we know the three components of the total annual cost, we can use a spreadsheet to determine the batch size that minimizes the total cost. Table 6.1 illustrates this for the data in Examples 6.1 to 6.4.

Observe that of the total annual cost *TC*, the order-cost component decreases when order size increases and the holding-cost component increases when order size increases. Figure 6.4 shows an optimal order size Q^* that minimizes total annual cost *TC*. *This optimal order quantity, Q^*, that minimizes total fixed and variable costs is called the* **economic order quantity (EOQ)**. From Table 6.1, we see that an order size of $Q = 3{,}000$ gives the total minimum cost. The EOQ can also be found analytically using calculus (see Appendix 6.1 for details) leading to a concise formula for the optimal order size, Q^*:

$$\textit{Optimal order size} = \sqrt{\frac{2 \times \textit{Fixed cost per order} \times \textit{Annual flow rate}}{\textit{Unit holding cost per year}}}$$

$$Q^* = \sqrt{\frac{2SR}{H}} \qquad \textbf{(Equation 6.5)}$$

popularly known as the **EOQ formula**.

Figure 6.4 shows all costs as functions of the order quantity as well as optimal order quantity and corresponding costs. Notice that the optimal order quantity exactly balances annual ordering and holding costs. Thus, we have, at optimality,

$$\begin{aligned}&\text{Total annual fixed costs per order} = \text{Total annual holding costs} \\ &S \times R/Q^* = H \times (Q^*/2)\end{aligned}$$

Table 6.1 Total Cost as a Function of Order Size: Spreadsheet Approach

Batch Size (Q)	Number of Orders (R/Q)	Annual Order Cost ($S \times R/Q$)	Average Cycle Inventory ($Q/2$)	Annual Holding Cost ($H \times Q/2$)	Annual Procurement Cost ($C \times R$)	Total Annual Costs (TC)
500	62.40	8,112.00	250	225.00	93,600	101,937.00
1000	31.20	4,056.00	500	450.00	93,600	98,106.00
1500	20.80	2,704.00	750	675.00	93,600	96,979.00
2000	15.60	2,028.00	1,000	900.00	93,600	96,528.00
2500	12.48	1,622.40	1,250	1,125.00	93,600	96,347.40
3000	10.40	1,352.00	1,500	1,350.00	93,600	96,302.00
3500	8.91	1,158.86	1,750	1,575.00	93,600	96,333.86
4000	7.80	1,014.00	2,000	1,800.00	93,600	96,414.00
4500	6.93	901.33	2,250	2,025.00	93,600	96,526.33
5000	6.24	811.20	2,500	2,250.00	93,600	96,661.20
5500	5.67	737.45	2,750	2,475.00	93,600	96,812.45
6000	5.20	676.00	3,000	2,700.00	93,600	96,976.00
6500	4.80	624.00	3,250	2,925.00	93,600	97,149.00

If we substitute for the optimal order quantity in the total annual cost expression given by Equation 6.4 and simplify, we find that the minimum annual total cost, TC^*, is

$$TC^* = \sqrt{2SRH} + CR \qquad \textbf{(Equation 6.6)}$$

where the first term ($\sqrt{2SRH}$) represents the total annual order and inventory holding cost at optimality and the second term (CR) is the total annual cost of materials.

Observe that to determine EOQ, we need to estimate three parameters, namely, the fixed cost per order S, the outflow rate R, and the unit holding cost per time H. In most practical settings, it is difficult to obtain accurate estimates of these parameters. A natural concern, then, is that the use of these parameter estimates in the EOQ formula

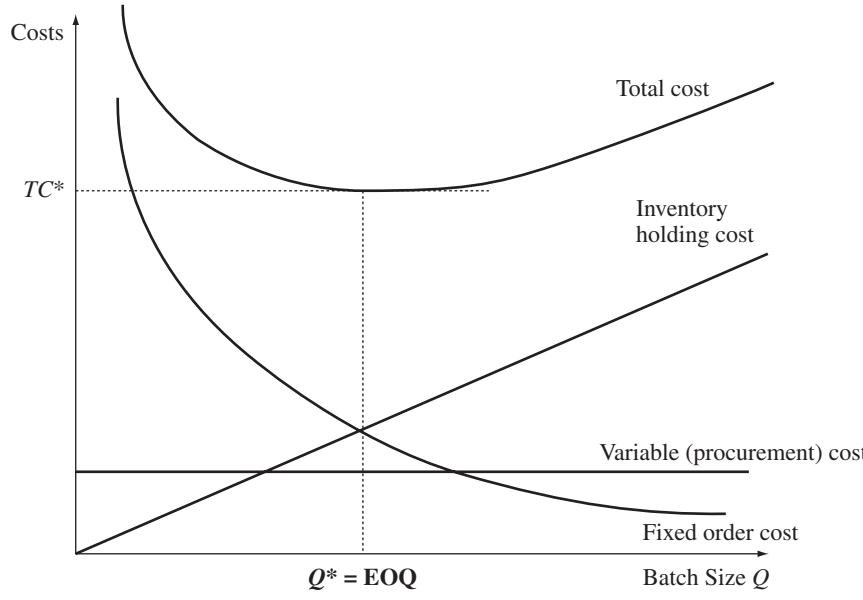


FIGURE 6.4 Total Annual Costs with Orders of Size Q

will not result in the truly optimal order quantity. While this is true, does it matter from a total cost perspective? From Figure 6.4, we observe that the total cost curve is relatively flat around EOQ; that is, some deviation from EOQ will not significantly increase total annual costs. Thus, even if the parameter estimates are not quite accurate, the total costs of the resulting policy will not deviate too far from the true optimal cost. Therefore, we say that the model is robust and practically useful. Similarly, even when the parameter estimates are accurate, a manager may wish to deviate from the resulting EOQ for considerations not included in this model. The robustness of the EOQ model guarantees that the cost consequences of such deviations will not be dramatic. Thus, EOQ provides a ballpark estimate of the range in which we should operate. Example 6.5 illustrates inventory management at a Centura hospital.

EXAMPLE 6.5

Using Examples 6.1 and 6.2, substituting known information into the EOQ formula yields

$$Q^* = \sqrt{\frac{2SR}{H}}$$
$$= \sqrt{\frac{2 \times \$130 \times 31,200}{\$0.90}} = 3,002$$

Thus, Centura should order IV Starter Kits in batches of 3,002 units whenever it places an order. The minor discrepancy with the spreadsheet approach of Table 6.1 arises as the spreadsheet approach evaluated costs for order quantities in steps of 500 instead of for each possible order quantity.

The resulting average cycle inventory will be

$$I_{cycle} = Q^*/2 = 1,501 \text{ units}$$

Using the Equation 6.4, we can calculate the minimum annual total cost TC^* as

$$TC^* = 130 \times (31,200/3,002) + 0.9 \times (3,002/2) + 3 \times 31,200 = \$96,302$$

This total results from $2,702 in ordering and inventory holding cost plus $93,600 in material cost. Average time spent by an IV Starter Kit in the input buffer can be computed as

$$T_i = I_{cycle}/R = 1,501/600 \text{ weeks} = 2.5 \text{ weeks}$$

Now, suppose that Centura's supplier prefers to ship in batch sizes of 3,500. It may be more convenient, therefore, to order 3,500 units at a time rather than the 3,002 specified by the EOQ formula. Deviating from the EOQ (in this case by 16.6%) increases total costs, but not much—if we substitute $Q = 3,500$ into the TC formula given by Equation 6.4, we find that the total annual cost would be only $31.86 higher than the minimum. This figure reflects an increase of 0.03% in total cost and 1.18% in order and inventory holding cost.

Three managerial insights follow from the EOQ formula and are discussed in the following sections.

Fixed Order Cost Reduction The optimal order size increases with the fixed order cost. The higher the fixed cost, the more we should order at a time in order to reduce the total number of orders per year. Conversely, lowering fixed cost would make ordering smaller quantities more economical which will reduce average inventory and flow time (see Example 6.6).

While discussion thus far has focused on decision making under economies of scale in procurement, it applies equally well when we have fixed costs in transportation

or production. Fixed costs in procurement usually include administrative costs of creating a purchase order, activities of receiving the order, and so on. Technology can be used to significantly reduce these costs. For example, creating electronic purchase orders takes less time and costs less. Firms that have adopted Electronic Data Interchange and the Internet for electronic purchasing have benefited from reduced fixed order costs, making it economical for them to order smaller quantities more frequently.

Fixed costs in transportation can be directly reduced by changing the mode of transportation—from a ship to a large truck (where possible) to a small truck to, perhaps, air. It can also be effectively reduced by sharing the fixed cost of a transportation mode across multiple supply or demand locations. These issues are discussed in more detail in supply chain management textbooks, such as Chopra and Meindl (2009).

Finally, reducing the setup or changeover costs can lower fixed costs in production. As we will see in Chapter 10, reducing setup times and hence costs has been a major factor in lean operations and just-in-time systems.

EXAMPLE 6.6

In Example 6.5, we have already seen, for instance, that ordering in batches of 3,002 results in a cycle inventory of 1,501 and adds 2.5 weeks to the flow time of IV Starter Kits through a Centura hospital. Now suppose that Centura wants to reduce the cycle inventory by half. In order to do so, it must reduce order size to 1,501—a change that would reduce flow time by 1.25 weeks. Since the optimal order size of 3,002 yields minimum total cost, any deviation from it without changing other factors will only increase the total cost. Recall, however, that one key lever available to Centura is reducing the fixed order cost *S*. Using the EOQ formula to solve for $S = Q^2H/2R$, we can infer that in order for 1,501 to be its optimal order size, Centura should reduce its fixed order cost *S* to \$32.50 (from the current value of \$130). Figure 6.5 gives the inventory profile for the base case as well as for the reduced fixed cost. Investment in information technology for procurement along with innovative ways to reduce fixed costs of transportation may be necessary to achieve this reduced level of fixed order cost.

Inventory versus Sales Growth From the EOQ formula, we observe that the optimal batch is proportional to the square root of outflow rate. Quadrupling the outflow rate, therefore, will only double EOQ and thus average inventory and average flow time in the buffer. Therefore, a doubling of a company's annual sales does *not* require a doubling of cycle inventories. That is, inventory growth should not track sales growth.

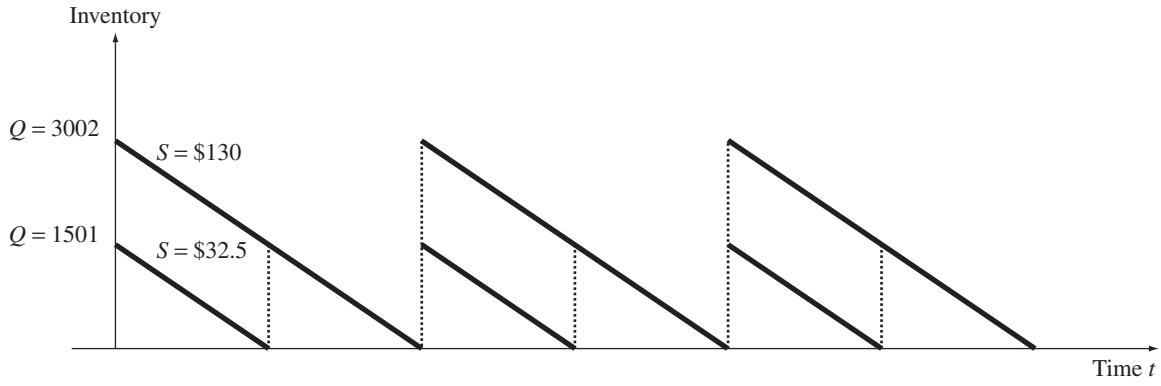


FIGURE 6.5 Inventory Profile with Reduction in Fixed Costs

Indeed, optimal inventory management would entail ordering more frequently, so that the 100 percent growth in throughput can be sustained by a mere 41 percent (from the square root of 2) increase in cycle inventory.

Centralization and Economies of Scale The fact that the optimal batch size is proportional to the square root of the outflow rate also leads to the idea of inventory centralization. For example, if a hospital network has multiple hospitals that order supplies independently, it can reduce its total cycle inventory by centralizing all the purchasing. For example, consider Centura Health introduced in Example 6.1. It has nine hospitals that order their supplies independently. Instead, Centura could centralize purchasing of all supplies and perhaps store these in a central warehouse. Under such a scenario, Centura will have to place orders for a total output flow rate that is nine times the output flow rate of each hospital. Assuming that the cost parameters remain unchanged in the central warehouse, we would expect the average inventory to be only three times (equal to the square root of 9) that of the decentralized hospital network. Example 6.7 illustrates the exact calculations.

EXAMPLE 6.7

As described in Example 6.1, Centura Health operates a network of nine hospitals. Currently, each hospital places orders with suppliers independently. Assume that each hospital operates under identical cost parameters (S = \$130 per order, and H = \$0.90 per unit per year) and that each satisfies a flow rate of 600 units per week. In Example 6.5, we computed the optimal order quantity for one of the hospitals as 3,002 units—the EOQ. The average cycle inventory of each hospital is then 3,002/2 = 1,501 units. Furthermore, the total annual order and holding cost for each hospital was \$2,702. If each hospital is assumed to be identical (in terms of the economics of placing orders and consumption of IV Starter Kits), the total cycle inventory across all nine hospitals is simply nine times the cycle inventory of each hospital operating independently and equals

$$9 \times 1{,}501 = 13{,}509 \text{ units}$$

with total annual order and holding costs as

$$9 \times \$2{,}702 = \$24{,}318$$

If Centura were to switch to purchasing via a central warehouse, then the total flow rate to be met from the new order process will be the total flow rate across all nine hospitals:

$$9 \times 31{,}200/\text{year} = 280{,}800 \text{ units/year}$$

Assuming that the cost parameters remain the same, the new EOQ is given by

$$\sqrt{\frac{2 \times \$130 \times 280{,}800}{\$0.90}} = 9{,}006$$

Corresponding average cycle inventory in the central

warehouse is equal to

$$\frac{9{,}006}{2} = 4{,}503$$

with a total annual order and holding costs of

$$\sqrt{2 \times 130 \times 280{,}800 \times 0.90} = \$8{,}106/\text{year}$$

which is 67% lower than for the decentralized operation.

Essentially, centralization gains advantage by exploiting economies of scale in placing orders. With increased volumes, it is economical to increase order size as well as order frequency; whereas each hospital operating independently placed about 10 orders per year, centralized purchasing entails ordering about 30 times a year. While the preceding discussion outlined a situation where the hospitals centralized purchasing *and* used a central warehouse, the latter is not essential. That is, the advantages of centralization can be achieved by simply centralizing the purchasing function. Under this scenario, each hospital will share its output flow rate information with the central coordinator. On consolidating the flow rates of each of the hospitals, the coordinator will place a single order with the supplier. The consolidated order can then be split and delivered to meet requirements of the respective hospitals. Obviously, such a practice will require capabilities in information technology and coordination.

6.6 EFFECT OF LEAD TIMES ON ORDERING DECISIONS

In many practical settings, process managers will have to make periodic ordering decisions. There are two fundamental questions that a process manager then needs to address:

1. How much to order?
2. When to reorder?

The first question depends on the trade-off between fixed costs of placing orders and the variable holding costs of carrying inventory resulting from ordering in quantities larger than one. An example of this essential trade-off was discussed in the previous section that led to the EOQ formula.

The second question depends on how long it takes to replenish inventory. *The time lag between the arrival of the replenishment and the time the order was placed is called the replenishment* **lead time**, which is denoted by L. Clearly, we should order at L units of time before we expect the inventory level to drop to zero.

Instead of keeping track of time, we can keep track of the inventory level and reorder as soon as the inventory drops below a certain reorder point, which is the available inventory at the time of placing an order. Such a policy, in which *a reorder is automatically triggered whenever inventory position reaches a specific limit is known as a* **continuous review policy**. We use ROP to denote the reorder point. Clearly, when we process continuously at a constant rate R, we should reorder when we have just enough inventory to cover requirements during the replenishment lead time L. Thus, the reorder point is found as

$$\text{Reorder point} = \text{Lead time} \times \text{Throughput}$$

$$ROP = L \times R \qquad \textbf{(Equation 6.7)}$$

If the lead time L is less than the time between orders (which we calculated earlier as $t_z = Q/R$ in Section 6.4), then the reorder point is the inventory that we have on hand at the time of placing an order. The reorder point decision can be superimposed on the inventory buildup diagram as shown in Figure 6.6. Example 6.8 illustrates the reorder point concept.

EXAMPLE 6.8

Recall that the replenishment lead time for ordering IV Starter Kits is $L = 1$ week. The reorder point is

$$ROP = L \times R = 1 \text{ week} \times 600 \text{ units/week} = 600 \text{ units}$$

Thus, whenever the input inventory level drops below 600 units, the process manager should place a new order with the supplier. Observe also that the 600 units is an in-transit inventory.

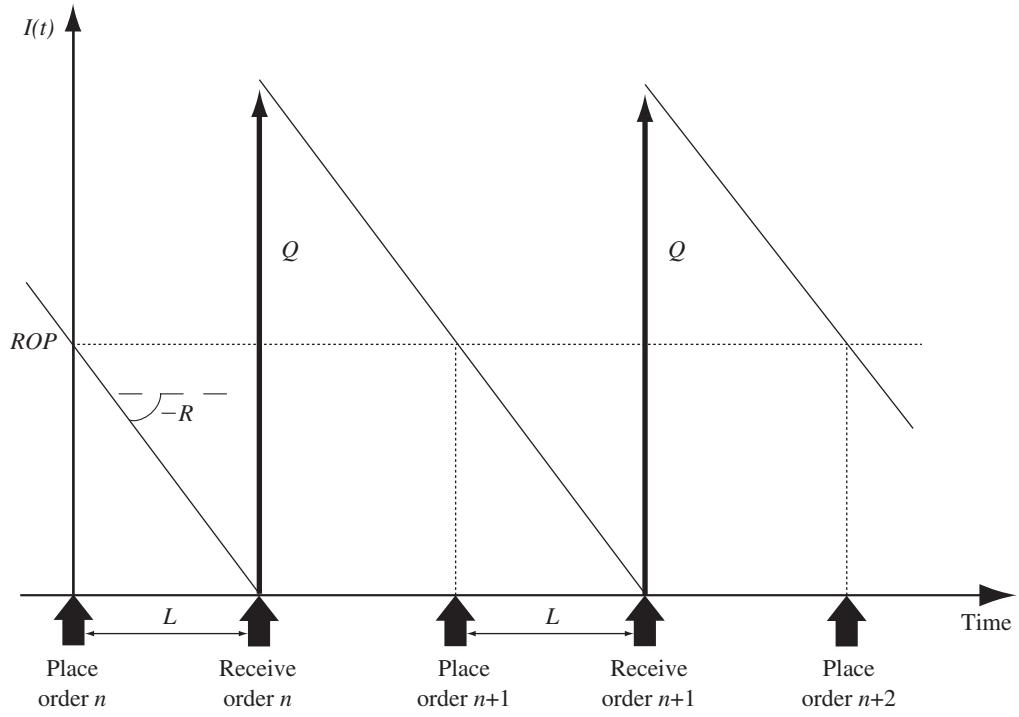


FIGURE 6.6 Ordering Decisions and the Reorder Point

If, however, the lead time L is larger than the time between orders (i.e., $L > Q/R$), the reorder point will be larger than the order quantity Q. This means that at the time we place our current order, there will be previous orders outstanding that will be received before the current order is received at a time L periods from now. In such cases, it is useful to define a measure called **inventory position** as

$$\textit{Inventory position} = \textit{Inventory level} + \textit{On-order inventory}$$

On-order inventory equals the sum of all outstanding orders. Observe that in Example 6.8, on-order inventory was zero and inventory level was equal to on-hand inventory. To articulate a reorder policy for the more general case, we then say that we place a reorder whenever the inventory position drops to the reorder point. Thus a **reorder point** is *the inventory position at the time a reorder is placed in a continuous review policy*. We illustrate by Example 6.9.

EXAMPLE 6.9

Suppose the replenishment lead time for ordering IV Starter Kits is $L = 6$ weeks (instead of the 1 week assumed in Examples 6.1 and 6.8). With the demand rate $R = 600$ units per week, the reorder point becomes

$$ROP = L \times R = 6 \text{ weeks} \times 600 \text{ units/week} = 3{,}600 \text{ units}$$

In Example 6.5, we calculated that the optimal order size is $Q^* = 3{,}002$ so that the time between ordering is $Q/R = 5.003$ weeks, which is less than the new lead time L of 6 weeks. Thus, there will always be one previous order outstanding at the time of placing the current order. Indeed, in this case, we say that we place a reorder whenever the inventory position reaches the reorder point of 3,600; an inventory position of 3,600 represents

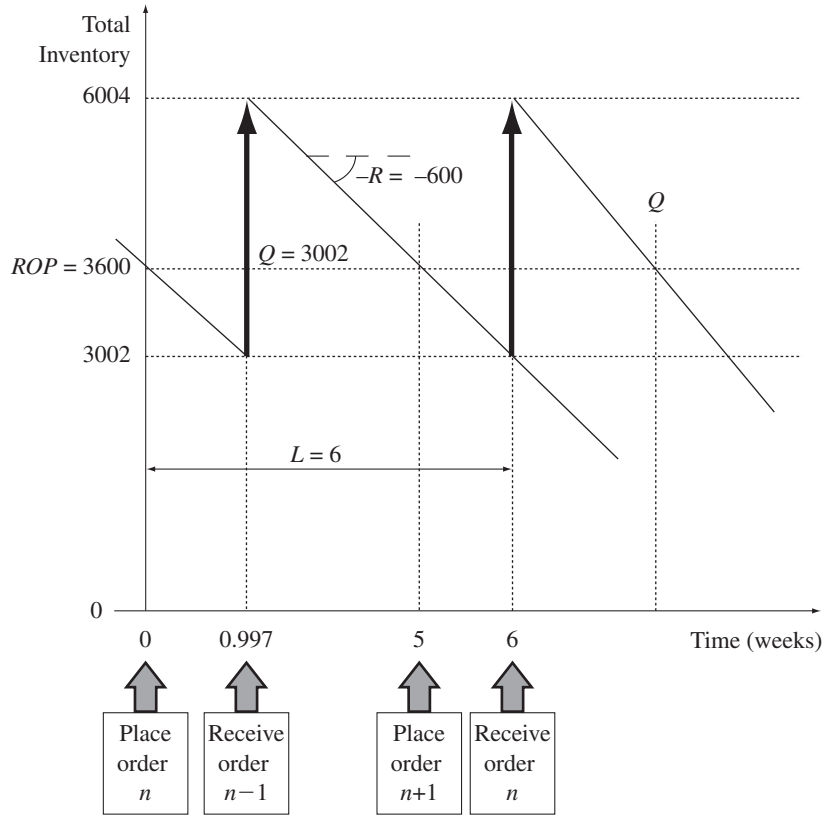


FIGURE 6.7 Ordering Decisions and the Re-order Point for Example 6.9

the sum of inventory level or on-hand inventory of 598 and one outstanding order ($Q^* = 3{,}002$) at the time of placing an order. The corresponding ordering decisions over time are shown in Figure 6.7.

So far, we have assumed that the output flow rate R and the lead time L are known with certainty. However, in reality this is rarely the case. For example, consumer demand is seldom known with certainty, and suppliers are not always reliable in their delivery schedules. We will see in the next chapter how to adjust the reorder point to incorporate a safety cushion, called safety inventory, to protect against this uncertainty.

6.7 PERIODIC ORDERING

So far we have illustrated the situation when a decision maker has some estimate of fixed costs (S) to determine optimal batch sizes. With constant flow rates, we have seen that the ordering decision exhibits a periodic behavior; for example, in Example 6.4 the buyer would order supplies of IV Starter Kits approximately every five weeks (= 31,200/6000). Furthermore, using a continuous review policy reorders are placed when inventory position drops to a level equal to the reorder point (ROP). Often, in practice, managers may not have any information on the fixed costs to estimate the EOQ; nor might the firm have the capability to continously monitor inventory to trigger a reorder. Alternately, standard practice may dictate a specific periodicity; for example, a firm may place orders every other Monday. What is the implication of such ordering policy on inventory levels? How does one then account for replenishment lead times?

Suppose that procurement policy at Centura Hospitals dictates that inventory be reviewed and orders be placed every other Monday. Then the time between subsequent

orders is 14 days. Because the weekly consumption rate is 600 units, the total consumption over a 14-day period will be 1200 units. Therefore, the hospital needs to have a 14-day supply or 1200 units on-hand. Every other week the hospital materials manager will place orders to bring the inventory level to 1200 units. *A policy under which the inventory position is reviewed at fixed time intervals and an order placed at that time, if necessary,* is called a **periodic review policy**. A specific instance of such a policy is called the **order upto policy** *in which orders are placed at periodic intervals to bring inventory position to a fixed level,* called the **order upto level** (OUL). *The fixed time interval for review* is known as the **review period**. We then say that for the example discussed, OUL is equal to 1200 units, which would be sufficient were the supplier to deliver the kits instantaneously.

How do we account for the more realistic scenario that entails a positive replenishment lead time? Recall that the replenishment leadtime of IV Started Kits was 1 week. Clearly an OUL of 1200 units will be consumed by the time the next review is due. An order placed at that moment will not arrive for another week. Therefore, the OUL needs to be adjusted to include sufficient inventory to cover demand during the lead time. In this case, OUL = 1200 + 600 = 1800 units. Of course, as before 600 units is in-process or pipeline inventory. Figure 6.8 shows an inventory profile diagram highlighting both the on-hand inventory as well as the inventory position (equal to the sum of on-hand and in-process inventory).

More generally, suppose that the inventory review period is denoted by T_r and replenishment leadtime is L. Then the order upto level (OUL) is given by

$$OUL = R \times (T_r + L) \qquad \textbf{(Equation 6.8)}$$

Since the average order quantity, $Q = R \times T_r$, the inventory fluctuates between zero and a maximum of Q leading to an average cycle inventory

$$I_{cycle} = Q/2 = (R \times T_r)/2$$

and the pipeline inventory is $R \times L$.

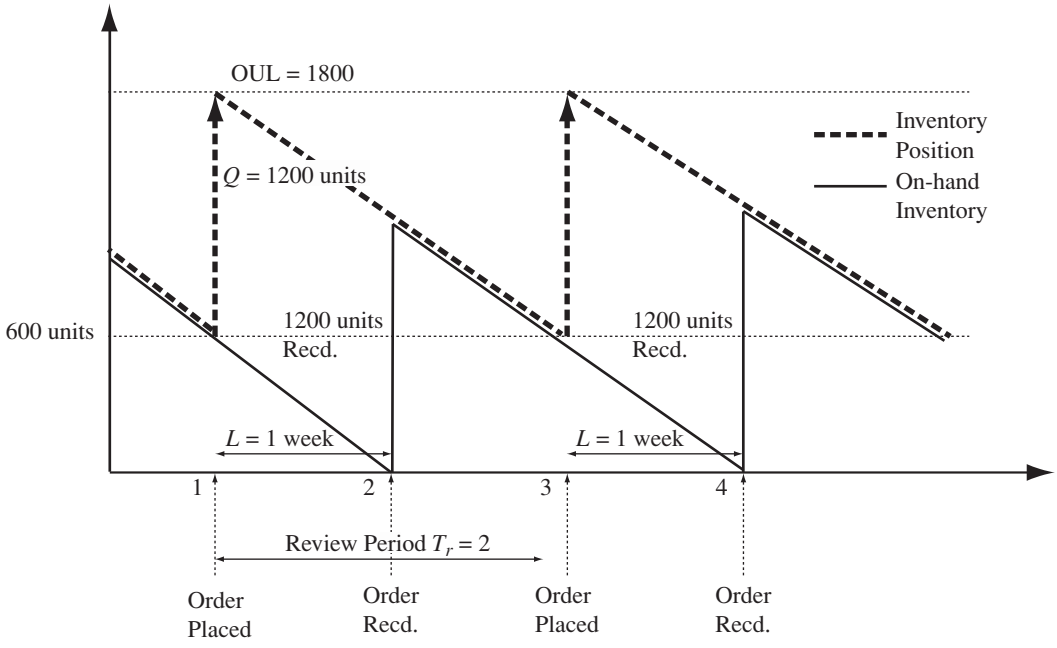


FIGURE 6.8 Inventory Profile for a Periodic Review Policy

For the example of Centura Hospitals above that follows a 14-day review period, the resulting cycle inventory will be 1200/2 = 600 units and pipeline inventory will be 600 units.

6.8 LEVERS FOR MANAGING INVENTORIES

We conclude by summarizing the most important ways of controlling the different types of inventories that we have discussed.

Theoretical Inventory Theoretical in-process inventory, expressed as

$$I_{th} = R \times T_{th}$$

is determined by throughput R and theoretical flow time T_{th}. As discussed in Chapters 4 and 5, managing these two measures can control inventory. T_{th} can be reduced by any one of the following measures:

- Reducing critical activity times
- Eliminating non-value-adding activities
- Moving work from critical to noncritical activities (as defined in Chapter 4)
- Redesigning the process to replace sequential with parallel processing

Reducing process flow rate R can also reduce theoretical in-process inventory. This option, however, will reduce the economic value of output per unit of time. Regardless, theoretical in-process inventory is usually only a small fraction of total inventory and managers like to reduce it primarily in order to reduce flow time.

Cycle Inventory Average cycle inventory is expressed as half the order size. Thus, cycle inventory can be reduced by reducing the order size. Recall that the optimal order size is given by the EOQ formula:

$$\textit{Optimal order size} = \sqrt{\frac{2 \times \textit{Fixed cost per order} \times \textit{Annual flow rate}}{\textit{Unit holding cost per year}}}$$

Alternately, when following a periodic policy, cycle inventory depends on the length of the review period.

Thus, the only sensible lever for reducing optimal cycle inventory (and hence the flow time) is to reduce the fixed order (or setup) cost or to reduce the review period. Simplifying the order process in conjunction with the use of information technology can reduce fixed order costs and review period. Investing in setup or changeover time reduction or investing in flexible resources helps lower fixed setup costs. Negotiating everyday low prices with suppliers instead of seeking short-time trade promotions can reduce excessive cycle inventories resulting from forward buying (see Appendix 6.2 for discussion).

Seasonal Inventory Seasonal inventory results from temporary fluctuations in outflows, coupled with the high costs of adjusting capacity to meet them. Using pricing and incentive tactics to promote stable demand patterns can reduce it. Increasing resource flexibility so that resources can process at various flow rates to match demand fluctuations will also make it less expensive to adjust seasonal inventory levels. Similarly, using flexible resources to produce countercyclical products makes it possible to level the load without having to build up inventory. A classic example is a company that produces snowblowers in winter and lawn mowers in summer, both with a single flexible production process.

Safety Inventory Safety inventory cushions the process against unexpected supply disruptions or surges in demand. The basic response to reducing its levels is reducing uncertainty in supply and demand. Ensuring reliable suppliers and stable demand

patterns largely eliminates the need for safety inventories. We will discuss the role of safety inventory more fully in Chapter 7.

Speculative Inventory Speculative inventory permits a firm to do one of two things:

1. Reduce the total cost of purchasing materials
2. Increase profits by taking advantage of uncertain fluctuations in a product's price

Negotiating stable prices, where possible, would eliminate speculative inventories and the associated portfolio risk.

Summary

Inventory accumulates whenever there is a mismatch between supply and demand. Along with flow time and flow rate, inventory is the third basic measure of process performance. In this chapter, we provided inventory classification, reasons and costs for holding inventory, optimal decisions under economies of scale, and short-term price promotions.

Depending on its location in the process, inventory can be classified as either inputs, in-process, or outputs inventory. Firms carry inventory for several reasons. First, a minimal level of inventory, called theoretical inventory, is needed to maintain a desired level of throughput in equilibrium. Transportation of products from one location to another involves delays; inventory being moved is classified as in-transit or pipeline inventory. To exploit economies of scale in production, procurement, and transportation, firms produce, purchase, or transport larger quantities than what may be immediately required, leading to cycle inventory. Faced with a seasonal demand and a desire to maintain a constant processing rate, firms create seasonal inventory. Firms may carry safety inventory to protect against demand or supply uncertainty. Finally, depending on the nature of price changes (e.g., random or promotional), firms may carry speculative inventory or forward buy more than what is needed. While there are several reasons for carrying inventory, it also entails a cost. Specifically, inventory carrying cost consists of physical holding costs as well as opportunity costs of capital tied up.

Decisions about purchasing under economies of scale involve a trade off between the fixed costs of ordering and the cost of holding the cycle inventory. As the lot size per order increases, fewer orders are placed in a year, reducing the annual fixed order costs. Increasing the lot size, however, increases cycle inventory, resulting in higher holding costs. The optimal lot size is determined by the economic order quantity formula. To reduce the cycle inventory, the lot size must be decreased. A primary lever to achieve this is to reduce the fixed costs of ordering or setup. Setup time reduction and using technology to cut purchase orders are some direct ways to reduce fixed order costs. In addition, aggregating purchases across multiple locations can also reduce lot sizes and hence cycle inventory. In particular, cycle inventory decreases by a factor of the square root of the number of locations aggregated.

In addition to the lot size, which determines the order quantity, a process manager needs to determine when to reorder, a decision that involves monitoring the inventory position, and placing an order when the inventory position drops to a reorder point. If demand is known perfectly, the reorder point is given by the demand during the lead time of replenishment. A policy under which reorders are automatically triggered when inventory reaches a certain level is known as a continuous review policy.

An alternate system of managing inventories is using a periodic review policy. Under such a policy, a process manager reviews inventory and places orders, if necessary, at fixed intervals. A specific instance of such a policy is the order upto policy. Order size, and hence the cycle inventory, depends on the length of the review period.

Finally, order quantities are also affected by price discount policies (see Appendix 6.2). Quantity discounts motivate the process manager to increase the order size. In response to short-term price reductions by a supplier, called a trade promotion, a buyer may order significantly more quantity than it normally does, leading to forward-buy inventory. "Everyday low pricing" is an effective tool to counter the buildup of forward-buy inventories.

Key Equations and Symbols

(Equation 6.1) $I_{th} = R \times T_{th}$

(Equation 6.2) $H = (h + r)C$

(Equation 6.3) $I_{cycle} = Q/2$

(Equation 6.4) $TC = S \times \frac{R}{Q} + H \times \frac{Q}{2} + C \times R$

(Equation 6.5) $Q^* = \sqrt{\frac{2SR}{H}}$

(Equation 6.6) $TC^* = \sqrt{2SRH} + CR$

(Equation 6.7) $ROP = L \times R$

(Equation 6.8) $OUL = R \times (T_r + L)$

where,
I_{th} = Theoretical inventory
R = Throughput or annual demand rate
T_{th} = Theoretical flow time
H = Total annual unit inventory holding cost
C = Unit variable cost
h = Unit physical holding cost as a fraction of unit variable cost
r = Unit opportunity cost of capital as a fraction of unit variable cost
I_{cycle} = Cycle inventory
Q = Order size
TC = Total annual cost
Q^* = Economic order quantity
TC^* = Total optimal annual cost
ROP = Reorder point
L = Replenishment lead time
T_r = Review period
OUL = Order upto level
S = Fixed cost per order

Key Terms

- Aggregate production planning
- All unit quantity discount policy
- Batch
- Batch size
- Chase demand strategy
- Continuous Review Policy
- Cycle inventory
- Economic order quantity (EOQ)
- Economies of scale
- EOQ formula
- Everyday low pricing (EDLP)
- Everyday low purchase prices (EDLPP)
- Fixed order cost
- Fixed setup cost
- Forward buying
- Incremental unit quantity discount policy
- In-process inventory
- Input inventory
- In-transit inventory
- Inventory holding cost
- Inventory level
- Inventory position
- Lead time
- Level-production strategy
- On-order inventory
- Opportunity cost
- Order upto level
- Order upto policy
- Output inventory
- Periodic review policy
- Physical holding cost
- Pipeline inventory
- Procurement batch
- Production batch
- Quantity discount policy
- Reorder point (ROP)
- Review period
- Safety inventories
- Safety stock
- Seasonal inventories
- Speculative inventory
- Theoretical inventory
- Trade promotion
- Transfer batch
- Work-in-process inventory

Discussion Questions

6.1 Explain how better inventory management affects a firm's bottom line.

6.2 Why do firms carry inventory even though it is costly to do so?

6.3 What are the key trade-offs in determining the economic order quantity?

6.4 Explain why it is not absolutely critical to estimate the cost parameters accurately in implementing the economic order quantity model.

6.5 Explain why fixed costs must decrease by a factor of four when reducing cycle inventory only by one half.

6.6 How can the use of information technology result in lower inventory?

6.7 Discuss whether reduction in replenishment lead times will reduce cycle inventory.

6.8 Which policy—continuous review or periodic review—results in a larger cycle inventory? Explain why.

Exercises

6.1 Suppose you purchase a part from a supplier for a unit cost of $4 with which you assemble red widgets. On average, you use 50,000 units of this part each year. Every time you order this particular part, you incur a sizable ordering cost of $800 regardless of the number of parts you order. Your cost of capital is 20% per year.

a. How many parts should you purchase each time you place an order?

b. To satisfy annual demand, how many times per year will you place orders for this part?

***6.2** BIM Computers Inc. sells its popular PC-PAL model to distributors at a price of $1,250 per unit. BIM's profit margin is 20%. Factory orders average 400 units per week. Currently, BIM works in a batch mode and produces a four-week supply in each batch. BIM's production process involves three stages:

- PC board assembly (the automatic insertion of parts and the manual loading, wave soldering, and laser bonding of electronic components purchased from outside sources)
- Final assembly
- Testing

When the firm wants to change production from one model to another, it must shut down its assembly line for half a day, which translates into four working hours. The company estimates that downtime costs half an hour of supervisory time and an additional $2,000 in lost production and wages paid to workers directly involved in changeover operations. Salaries for supervisory personnel involved amount to $1,500 per day.

Although BIM products are generally regarded as high quality, intense price competition in the industry has forced the firm to embark on a cost-cutting and productivity improvement campaign. In particular, BIM wants to operate with leaner inventories without sacrificing customer service. Releasing some of the funds tied up in outputs inventory would allow BIM to invest in a new product development project that is expected to yield a risk-adjusted return of 20% per annum. Assume 50 workweeks in a year and five working days in a week.

a. Determine BIM's total annual cost of production and inventory control.

b. Compute the economic batch size and the resulting cost savings.

6.3 Victor sells a line of upscale evening dresses in his boutique. He charges $300 per dress, and sales average 30 dresses per week. Currently, Victor orders a 10-week supply at a time from the manufacturer. He pays $150 per dress, and it takes two weeks to receive each delivery. Victor estimates his administrative cost of placing each order at $225. Because he estimates his cost of inventory at 20%, each dollar's worth of idle inventory costs him $0.30 per year.

a. Compute Victor's total annual cost of ordering and carrying inventory.

b. If he wishes to minimize his annual cost, when and how much should Victor order in each batch? What will be his annual cost?

c. Compare the number of inventory turns under the current and proposed policies.

***6.4** A retailer estimates her fixed cost for placing an order at $1,000. Currently, she orders in optimal quantities of 400 units. She has, however, heard of the benefits of just-in-time purchasing—a principle that advocates purchasing goods in smaller lots as a means of keeping inventory down. To do so, she needs to reduce her fixed order costs. What should her fixed ordering costs be if she wishes her order size to be no larger than 50?

6.5 Major Airlines would like to train new flight attendants in an economically rational way. The airline requires a staff of about 1,000 trained attendants to maintain in-flight service. Because of the nature of the job, attendants have a high propensity to quit, with average job tenure being about two years, hence the need to train new attendants. Major's training course takes six weeks, after which trainees take one week of vacation and travel time before entering the pool from which they are assigned to flight duty as needed to fill vacancies created by attrition. To reduce the dropout rate and ensure the continued availability of trained attendants, Major pays trainees $500 per month while they are training, vacationing, and waiting for assignment.

a. The cost of the training itself consists mainly of salaries for instructors ($220 per person per week) and support personnel ($80 per person per week). A training team consists of 10 instructors and 10 supporting personnel. The team is paid only for the time engaged in training, and pay is independent of both class size and the number of classes running simultaneously. Assume 50 work weeks in a year. Determine the most economical size of a trainee class, the annual total cost of this policy, and the time interval between starting consecutive classes. Draw the inventory-time diagram, showing when each batch will begin and end training, when each will take vacation time, and when each will be ready for duty.

b. Now modify the solution obtained in part (a) so that only one class will be in training at one time. Note that this requirement means that a new class must start every six weeks. Determine the corresponding class size and the total annual cost of this operation. Compare your findings for this option with the optimum cost for the option described in part (a) and make a recommendation as to which option Major Airlines should choose.

6.6 National Bank operates a network of automated teller machines (ATMs). Cash withdrawals at an ATM average about $80. The bank estimates that the fixed cost of filling an ATM with cash is about $100 regardless of the amount of cash put in. Average numbers of cash withdrawals per week is about 150. How much cash should the bank keep at an ATM if its annual cost of money is 10%? How often should the bank replenish an ATM?

6.7 (See Appendix 6.2) Gourmet Coffee (GC) is a specialty coffee shop that sells roasted coffee beans. It buys green beans, roasts them in its shop, and then sells them to the consumer. GC estimates that it sells about 150,000 pounds of coffee per year. Green beans cost about $1.50 per pound. In addition, there is a shipping charge that GC pays its supplier according to the following schedule:

Quantity Shipped	Shipping Cost per Pound
Less than 10,000 pounds	$0.17
Less than 15,000 pounds	$0.15
More than 15,000 pounds	$0.13

GC estimates its cost of inventory at 15% per year. The administrative cost of placing an order (fax/phone/billing) and receiving the goods and so on is about $50 per order. In addition, to receive a shipment into its shop, GC rents a forklift truck for $350.

a. What is the optimal order quantity of beans for GC? What is the total annual cost?

b. GC is considering buying a forklift and building a ramp that will allow it to eliminate the rental cost of a forklift. GC will have to borrow money to finance this investment. If the life of the equipment is approximately five years, how much money should GC be willing to spend to buy a forklift and build a ramp? If the investment were made, what should be the optimal order policy for GC?

6.8 A supplier to Toyota stamps out parts using a press. Changing a part type requires the supplier to change the die on the press. This changeover currently takes four hours. The supplier estimates that each hour spent on the changeover costs $250. Demand for parts is 1,000 per month. Each part costs the supplier $100, and the supplier incurs an annual holding cost of 25%.

a. Determine the optimal production batch size for the supplier.

b. Toyota would like the supplier to reduce their batch size by a factor of four; that is, if the supplier currently produces Q parts per batch, Toyota would like them to produce $Q/4$ parts per batch. What should the supplier do in order to achieve this result?

6.9 Superfast Motors manufactures and sells a wide variety of motors to industrial customers. All motors cost about the same and are assembled on the same line. Switching over from assembling one motor to another requires about two hours. Superfast assembles motors to be stocked in a distribution center from where they are shipped as orders arrive. HP is the highest-selling motor (in terms of units sold) and LP the lowest selling.

a. Will the average cycle inventory of HP motors be:
- Higher than the cycle inventory of LP motors
- Lower than the cycle inventory of LP motors
- Same as the cycle inventory of LP motors?

b. Will the average time spent by an HP motor in inventory be:
- Higher than the time spent by an LP motor
- Lower than the time spent by an LP motor
- Same as the time spent by an LP motor?

***6.10** Complete Computer (CC) is a retailer of computer equipment in Minneapolis with four retail outlets. Currently each outlet manages its ordering independently. Demand at each retail outlet averages 4,000 units per week. Each unit costs $200, and CC has a holding cost of 20% per annum. The fixed cost of each order (administrative plus transportation) is $900. Assume 50 weeks in a year.

a. Given that each outlet orders independently and gets its own delivery, determine the optimal order size at each outlet.

b. CC is thinking of centralizing purchasing (for all four outlets). In this setting, CC will place a single order (for all outlets) with the supplier. The supplier will deliver the order on a common truck to a transit point. Since individual requirements are identical across outlets, the total order is split equally and shipped to the retailers from this transit point. This entire operation has increased the fixed cost of placing an order to $1,800. If CC manages ordering optimally in the new setting, compute the average inventory in the CC system (across all four outlets).

Selected Bibliography

Buzzell, R. D., J. A. Quelch, and W. J. Salmon. "The Costly Bargain of Trade Promotions." *Harvard Business Review* (March–April 1990): 1–9.

Chopra, S., and P. Meindl. *Supply Chain Management: Strategy Planning and Operations.* 4th ed. Upper Saddle River, N.J.: Prentice Hall, 2009.

Hadley, G., and T. M. Whitin. *Analysis of Inventory Systems.* Upper Saddle River, N.J.: Prentice Hall, 1963.

Nahmias, S. *Production and Operations Analysis.* 6th edition. Homewood, Ill.: McGraw-Hill/Irwin, 2008.

Peterson, R., and E. A. Silver. *Decision Systems for Inventory Management and Production Planning.* New York: John Wiley & Sons, 1979.

Sasser, W. "Match Supply and Demand in Service Industries." *Harvard Business Review* (November–December 1976): 132–138.

APPENDIX **6.1**

Derivation of EOQ Formula

DERIVATION OF EOQ FORMULA

From Equation 6.4 the total annual costs is given by

$$TC = S \times \frac{R}{Q} + H \times \frac{Q}{2} + C \times R$$

Taking the first derivative of the total cost function TC with respect to Q yields

$$d(TC)/dQ = -SR/Q^2 + H/2$$

If we set the first derivative of the total cost function equal to zero (which is a condition to minimize TC), solving for Q yields the EOQ formula as

$$Q^* = \sqrt{\frac{2SR}{H}}$$

APPENDIX **6.2**

Price Discounts

In addition to fixed order costs, scale economies in procurement can be driven by price discounts that a supplier may offer to a buyer who purchases in large quantities. Consider the situations described in Examples 6.10 and 6.11.

EXAMPLE 6.10

The supplier of IV Starter Kits offers the following price schedule to Centura Hospitals:

Order Quantity	Unit Price
0–5,000	$3.00
>5,000–10,000	$2.96
>10,000	$2.92

Should the buyer alter its purchasing decision determined in Example 6.4?

EXAMPLE 6.11

A buyer at the discount retailer, Target is considering ordering Colgate toothpaste for its stores. Demand for Colgate toothpaste is estimated to be 10,000 tubes per month. The fixed cost of an order—including administrative, transportation, and receiving—is estimated to be $100. The unit annual holding cost is 20%. The regular unit purchase price is $3. The manufacturer offers a one-time discount of 5% for units purchased over the next one month.

Price discounts take many forms. *A policy where prices depend on the quantity purchased* is known as a **quantity discount policy**. A commonly used quantity discount policy known as the **all unit quantity discount policy** where *a buyer receives discount on all units purchased whenever the quantity purchased exceed a certain threshold*. The pricing policy described in Example 6.10 is an all-unit quantity discount policy because the reduced price of $2.96 applies to all units above 5000. In contrast, under what is also known as an **incremental quantity discount policy**, *a buyer receives discount only on additional units purchased above a certain threshold value and charges the regular price for units upto the threshold*. As illustration, suppose the incremental discount prices and thresholds for price breaks were similar to those in Example 6.11. If the buyer places an order of 5500 units, then whereas under the all-unit discount scheme the purchase price for all 5500 units is $2.96, under the incremental quantity discount scheme, the price for the first 5000 units is $3.00 and the remaining 500 units are charged $2.96. The EOQ formula needs to be modified to accommodate quantity discounts; we skip the details of EOQ models for quantity discount pricing and refer the reader to inventory or supply-chain management texts such as Chopra and Meindl (2009) and Nahmias (2008). However, a spreadsheet approach, similar to the one described in Table 6.1 for an undiscounted case, can always be used to find the optimal order quantity.

The example in Example 6.11 is a short-term discount policy where discounts are offered for only a short period of time, known as a trade promotion. The supplier offers incentives in the form of one-time opportunities to sell materials at reduced unit costs or perhaps notifies the buyer of an upcoming price increase and offers one last chance to order at the lower price. In both cases, of course, the buyer has an incentive to fill future needs by purchasing a single large quantity at the reduced price. *Taking advantage of such an opportunity by purchasing for future needs today* is called **forward buying**.

Short-term trade promotions could motivate a retailer to forward buy large quantities of material, resulting in a substantial increase in inventory. According to a study of food distributors by Buzzell et al. (1990), forward-buy inventories normally amounted to 40% to 50% of total stocks. Of course,

the total costs of trade promotions and resulting forward buy includes, in addition, added transportation and handling costs, higher administrative and selling costs that both suppliers and distributors incur, and costs of time buyers spend trying to evaluate deals. These added costs could be substantial. For example, Buzzell et al. (1990) report that the cost of forward buys in the nonperishable food-store products account for at least 1.15% to 2.0% of retail sales.

It can be shown that order increases designed to take advantage of short-term discounts can generate significant increases in inventory, and thus material flow time, in the supply chain (Chopra and Meindl, 2009). This realization has led many firms to adopt a policy of **everyday low pricing (EDLP)**—*a pricing policy whereby retailers charge constant, everyday low prices with no temporary discounts*. With EDLP, customers will not exercise forward buying. The same argument can be used upstream in the supply chain. If wholesalers practice **everyday low purchase prices (EDLPP)**, charging constant prices with no discounts, retailers will not forward buy. Thus, flows in the entire chain will be smoother and total inventories lower than when forward buying is practiced. We will examine the implications of such policies for flows in supply chain management in Chapter 10.

This page intentionally left blank

PART III

Process Flow Variability

Chapter

7

Managing Flow Variability: Safety Inventory

INTRODUCTION

In the 1990s, General Electric (GE) Lighting served its European customers through a distribution network that consisted of seven warehouses, including three near Paris and one each in Austria, Belgium, Germany, and Switzerland. The network of multiple warehouses was built on the premise that it will allow GE Lighting to be "close to the customer." Contrary to expectations, establishing the distribution network led to an "inventory-service crisis." Inventory levels in the network were high and customer service suffered. GE Lighting wanted to reevaluate its distribution strategy in Europe while also expanding to serve southern Europe. They faced several questions. What are the key drivers of inventory when customer demands are unpredictable? Should they invest in a better forecasting system? What should be the right inventory level? What service level is appropriate to offer? Should they continue to serve their customers using a decentralized network or build a central distribution facility to serve all customers?

GE Lighting ultimately consolidated the original seven warehouses into a single master distribution center in France to serve customers in Austria, Belgium, France, Germany, the Netherlands, Luxembourg, and Switzerland. In addition, to serve customers in other parts of Europe, it opened a facility in Sweden to serve the Scandinavian customers, one

each in the United Kingdom and Italy to serve customers in those countries, and a distribution center in Spain to serve customers in both Spain and Portugal (Harps, 2000).

Matching inflows (supply) and outflows (demand) is a critical aspect of managing any business process. In Chapter 6, we focused on economies of scale to explain why firms may plan supply in excess of demand and hold the resulting inventory. Actual supply may still fall short of demand because of unpredictable variability (uncertainty) in either supply or demand. This may result in process starvation, product shortages, lost sales, and customer dissatisfaction. Several companies find themselves in this perilous situation, often with severe financial or nonfinancial consequences, as illustrated in Table 7.1. The process manager may respond by holding additional inventory—called safety inventory—as a cushion, or buffer, that absorbs fluctuations, thus maintaining stock availability despite variability in supply or demand.

In this chapter, we explore this protective function of inventories, its key determinants, and the managerial levers available to control these inventories. As in Chapter 6, our discussion applies equally to buffers at any one of the three stages in a process: input (raw material), in process, and output (finished goods). For consistency, however, we refer to inflows into the buffer as supply and outflows from the buffer as demand.

To plan an adequate level of inventory, the process manager needs to forecast demand. The amount of safety inventory required will then depend on the accuracy of that forecast. In Section 7.1, we outline some general principles about forecasts that bear on the management of safety inventory. The rest of the chapter then examines these implications in greater detail. In Section 7.2, we begin by studying the amount of stockout protection provided by a given level of inventory and the amount of

Table 7.1 Examples of Supply–Demand Mismatch

Apple's iPhone broke sales record when it sold 1.7 million units on release day. Yet people were lining up to buy the gadget a week later. It is estimated that Apple could have sold up to 2 to 2.5 million if could produce more units.

Financial Times, January 2011

During 2007, Ninentdo's game system Wii was hard to get due to supply shortages. Analysts estimate that the company was leaving close to $1.3 billion on the table in unmet demand.

techspot.com, December 17, 2007

Mumbai's real estate is said to be hot property. However, in the last quarter, sales have dipped so low that builders are getting worried. . . . At the current pace of consumption, it will take two years and four months to exhaust this stock. This is alarming because, a healthy market is supposed to have only an eight-month inventory pile-up.

MumbaiMirror.com, February 8, 2011

An inventory write-off widened fourth quarter losses at Bluefly, despite a substantial increase in revenues at the online fashion retailer. Fourth quarter revenues were up 10 percent to US$29.7 million, but the inventory write-off knocked back gross profit by 7 percent, while the company's net loss for the quarter widened to $5.6 million from $3.5 million last year.

Just-Style.com, March 27, 2008

In a December report released by the Canadian Pharmacists Association, nearly 90 percent of pharmacists across the country said shortages have greatly increased in the past year. Antibiotics, antinausea, and heart drugs are among the top medications that pharmacists say are in shortest supply . . . people who can't get access to their primary drug of choice, may be forced to go without or take alternatives, which could lead to serious side effects . . . left unabated, the situation could cause someone with depression to commit suicide or lead other patients to experience serious health problems because they couldn't get the drugs they needed.

The Globe and Mail, January 31, 2011

safety inventory required to provide a given level of protection. In Section 7.3, we consider the problem of determining the optimal level of protection that balances the expected costs of overstocking and understocking. Section 7.4 examines the factors affecting variability in supply and demand and thus the extent of safety inventory needed to provide certain levels of service. Sections 7.5 and 7.6 outline operational strategies for reducing variability by means of aggregation of demand and postponement of supply. Section 7.7 illustrates the periodic review order policy. Finally, Section 7.8 summarizes the key levers for managing safety inventory and customer service in the face of demand variability.

7.1 DEMAND FORECASTS AND FORECAST ERRORS

Until now (e.g., discussions in Chapter 6), we have assumed that product demand is known and is constant over time. In reality, of course, demand usually varies over time. Although some variation is systematic and hence predictable (e.g., because of trends or seasonality), much of it results from unpredictable, unexplainable, random factors called noise. As *a process of predicting future demand,* **forecasting** is, among other things, an effort to deal with noise. Firms forecast a variety of factors, such as future customer demand, sales, resource requirements and availabilities, and interest rates.

Forecasting Methods A variety of forecast methods are available; they can be classified broadly as *subjective* or *objective.* Subjective methods are based on judgment and experience and include customer surveys and expert judgments. Objective methods are based on data analysis. The two primary objective methods are causal models and time-series analysis. **Causal models** are *forecasting methods that assume that in addition to the data, other factors influence demand.* For example, future sales could be a function of consumer prices. **Time-series analyses** are *forecasting methods that rely solely on past data.* Objective methods aim to filter out noise and estimate the effect of such systematic components as trends and patterns of seasonality or such causal factors as the effect of price on sales.

A detailed discussion of forecasting methods is beyond the scope of this book, but they are discussed in Chopra and Meindl (2009) and Nahmias (2008). Our focus in this section will be on some general characteristics of forecasts, as identified by Nahmias (2008), that process managers should understand—regardless of the forecasting method that they may use—to make rational decisions about process inventory. These general characteristics are the following:

1. ***Forecasts are usually wrong:*** Even if we could accurately estimate variations in the systematic components of a demand pattern, the presence of random noise that we can neither explain nor control leads to inaccuracy. Therefore, decisions made on the basis of a forecast (specified as a single number) could have unexpected consequences in terms of either higher costs or inadequate service.
2. ***Forecasts should, therefore, be accompanied by a measure of forecast error:*** A measure of forecast error quantifies the process manager's degree of confidence in the forecast. Our decisions (e.g., regarding inventory) should change with our confidence in the forecast—the greater the forecast error, the greater the chance of a stockout for a given level of safety inventory. We will study the exact relationship between the safety inventory, the service level, and the forecast error in Section 7.2.
3. ***Aggregate forecasts are more accurate than individual forecasts:*** For example, forecasting demand for sweaters by individual colors is less reliable than forecasting total demand for all sweaters. Intuitively, we know that aggregation reduces variability—or, more precisely, reduces the amount of variability relative to

aggregate mean demand. Why? High- and low-demand patterns among individual products tend to cancel one another, thereby yielding a more stable pattern of total demand. As a result, less safety inventory is needed in the aggregate. This realization underlies the principle of reducing variability and safety inventory by pooling and centralizing stock—which we will discuss in Section 7.5.

4. ***Long-range forecasts are less accurate than short-range forecasts:*** Again, intuitively we know that events further in the future are less predictable than those that are more imminent. Every meteorologist knows that forecasting tomorrow's weather is easier than forecasting next week's weather. Likewise, matching supply and demand in the short run is easier than planning for the long term. The closer to the actual time of demand a manager can make supply decisions, the more information will be available to make those decisions. Short-range demand forecasts, therefore, will be more accurate than long-range demand forecasts, and less safety inventory will be needed. Section 7.6 focuses on the use of postponement strategies to exploit short-range forecasts.

In addition to incorporating hard quantitative data, forecasts should be modified to include qualitative factors such as managerial judgment, intuition, and market savvy. After all, forecasting is as much an art as a science, and no information should be ignored.

7.2 SAFETY INVENTORY AND SERVICE LEVEL

If we grant that forecasts are usually wrong, we must also agree that planning supplies so that they merely match demand forecasts will invariably result in either excess inventories when supply exceeds demand or stockouts when demand exceeds supply, as illustrated in Example 7.1.

EXAMPLE 7.1

Consider a GE Lighting warehouse near Paris and the procurement decisions faced by the warehouse manager for its industrial flood lamp. The throughput rate of lamps is, say, 2,000 units per day.[1] The warehouse manager orders a batch of 28,000 lamps, equivalent to a 14-day supply. Whenever the manager places an order, the replenishment is received in 10 days. The manager reorders whenever the inventory level drops to 20,000 units. He estimates that the cost of holding one lamp in inventory for one year is €20.

How was the throughput rate of 2,000 units per day established? It was perhaps based on some forecast of the number of lamps demanded, but the forecast inevitably will involve some error. Observe that the manager has set the reorder point to 20,000 units and the replenishment lead time is 10 days. During that 10-day leadtime, one of the following events will inevitably occur:

1. Actual requirements will fall below 20,000 units, resulting in excess inventory.
2. Actual requirements will exceed 20,000 units, resulting in a lamp stockout.

Only by extreme coincidence will actual demand be *exactly* 20,000 units. If demand is equally likely to be above or below 20,000, then there is a 50% probability that keeping an inventory of 20,000 units will result in a stockout.

Stockouts occur whenever demand exceeds supply; they have critical business implications. In the GE Lighting's Paris warehouse situation, lamp stockouts imply that

[1]All numbers in the examples are fictitious and used only to illustrate the concepts.

customer demands will go unsatisfied. That may mean lost customers and lost revenue as well as loss of customer goodwill, which may lead to lost future sales. A comprehensive study on out-of-stock frequency in the retail sector (Gruen et al., 2002) has estimated that worldwide the out-of-stock frequency in these settings averages at 8.3 percent. The researchers estimate that a typical retailer loses about 4 percent of sales because of having items out of stock. A 4 percent loss of sales for the average firm in the grocery retail sector, for example, translates into earnings-per-share loss of 4.8 percent.

Sometimes, *customers may be willing to wait and have their needs satisfied later, in which case their demand is said to be* **backlogged**. Regardless, when a stockout occurs, customer needs are not immediately fulfilled, and this leads to some level of dissatisfaction. To avoid stockouts—and to provide better customer service—businesses often find it wise to keep extra inventory just in case actual demand exceeds the forecast. As mentioned earlier, *inventory in excess of the average or in excess of forecast demand* is called **safety inventory** or **safety stock**.

This definition of safety stock may seem to imply that it is always *positive.* Depending on costs and benefits of carrying inventory, however, it may be preferable to keep an inventory level that covers less-than-average demand, which yields a negative safety inventory. We will explore negative safety inventory in Section 7.3.

7.2.1 Service Level Measures

To determine the optimal level of safety inventory, the process manager should consider economic trade-offs between the cost of stockouts and the cost of carrying excess inventory. Although inventory-carrying costs are quantifiable (as identified in Chapter 6), unfortunately the intangible consequences of stockouts are difficult to evaluate in monetary terms. Consequently, the process manager often decides to provide a certain level of customer service and then determines the amount of safety inventory needed to meet that objective. The two commonly used measures of customer service are as follows:

- **Cycle service level** refers to either *the probability that there will be no stockout within a time interval* or, equivalently, the proportion of time intervals without a stockout, where the time interval of interest will depend on the type of inventory control policy used (to be elaborated later).
- **Fill rate** is *the fraction of total demand satisfied from inventory on hand.*

These measures are illustrated in Example 7.2.

EXAMPLE 7.2

Suppose that a process manager observes that within 100 time intervals, stockouts occur in 20. Cycle service level is then

$$80/100 = 0.8\text{, or } 80\%$$

That is, the probability of being in-stock is 80%. Now suppose that in each time interval in which a stockout occurred, we measure the extent of the stockout in terms of the number of units by which we were short. Specifically, suppose that cumulative demand during the 100 time intervals was 15,000 units and the total number of units short in the 20 intervals with stockouts was 1,500 units. The fill rate, therefore, is

$$1 - 1{,}500/15{,}000 = 13{,}500/15{,}000 = 0.9\text{, or } 90\%$$

In general, we can write the following expression for fill rate:

$$\begin{aligned}\text{Fill Rate} &= \text{Expected Sales}/\text{Expected Demand}\\ &= 1 - \text{Expected Stockout}/\text{Expected Demand}\end{aligned}$$

Whether or not an 80% cycle service level or a 90% fill rate is an acceptable measure of service will depend on several factors including product category, business context, competitive environment, etc.

Effective inventory policies can be devised to achieve a desired level of either measure of customer service. In most business-to-consumer transaction settings (e.g., retail sales), only information on sales is available, as true demand is rarely observed because of stockouts. This makes it difficult to measure fill rate, which requires knowledge of demand. Furthermore, analyzing inventory policies for cycle service level is often simpler than for the fill rate measure. In this book, we focus on cycle service level and refer to it simply as service level (*SL*). Discussions about inventory policies for the fill rate measure can be found in a supply chain management text (see, Chopra & Meindl, 2009).

In the rest of this section, we determine two items:

1. The service level provided by a given amount of safety inventory
2. The amount of safety inventory needed to provide a given level of service

Before we address these issues, we will describe a modification of the inventory policy (introduced in Chapter 6, Section 6.7) when demands are uncertain.

7.2.2 Continuous Review, Reorder Point System

As discussed in Chapter 6, in establishing an inventory system, a process manager must first decide how often the inventory level should be reviewed. The two choices are either reviewing it continuously (real time) or periodically (weekly, monthly). Obviously, the decision will depend on the cost of the review. With the widespread use of information systems, this cost has been declining—making for a compelling case to adopt a continuous review system. We first discuss inventory policy for a continuous review system. We also illustrate advanced concepts in management of inventory under demand uncertainty using the continuous review system. Later in Section 7.7, we discuss the implications of following a periodic review system.

Recall the two fundamental questions that a process manager must address once a review policy has been set:

- How much should I order?
- When should I reorder?

The answer to the first question depends on the trade-off between the fixed cost of placing orders and the variable holding cost of carrying the inventory that results from larger order quantities. This trade-off is essentially what led to the development of the economic order quantity (EOQ) formula discussed in Chapter 6. *Having initially ordered a fixed quantity, the process manager monitors inventory level continuously and then reorders (a quantity perhaps equal to EOQ) once available inventory position falls to a prespecified reorder point. This order policy, known as a* **continuous review**, **reorder point policy**, is essentially the one described in Chapter 6. Here we extend it to include uncertainty in demand and replenishment lead time.

In the sequel, we will use **boldface** notation to signify that the variable represented by the notation can take values that are uncertain or unknown. For example, **X** will be a variable that takes uncertain values; **X** is usually referred to as a random variable. The average value of **X** will be represented by an *italicized X*, and its standard deviation (a statistical measure of the variability of **X**; see Appendix II for details) will be represented by the symbol sigma with a subscript *X*, that is, σ_X.

In this context, **R** will denote the (uncertain) demand rate per unit of time (day, week, month, or year). The average demand rate is *R*, which now represents the average rate at which inventory is depleted over time. Actual demand rate—and thus inventory

level—will vary. Similarly, the (uncertain) replenishment lead time is denoted by **L** with an average value denoted by *L.* This delay can result from a combination of various delays in information, processing, or transportation. The variable is measured in the same time units (days, weeks, months, or years) as **R**. If **R** is the number of flow units demanded per day, week, or month, then **L** is measured in the number of days, weeks, or months, respectively, that elapsed between the placing of an order and its receipt. Thus, when the available inventory level falls to the reorder point, a new order of size Q (a quantity perhaps equal to EOQ) is placed that arrives in **L** time periods. On receipt of this new order, of course, available inventory level increases by Q units.

Leadtime Demand The reorder point inventory is used to meet flow-unit requirements until the new order is received **L** periods later. The risk of stockout occurs during this period of replenishment lead time. The *total flow-unit requirement during replenishment lead time* is called **Leadtime demand** and is designated by **LTD**. In general, if either flow rate **R** or leadtime **L** is uncertain, total leadtime demand **LTD** will also be uncertain. Uncertainty in flow rate results from less-than-perfect forecasting (which is inevitable). Uncertainty in leadtime may be due to a supplier's unreliability in delivering on-time orders. When the leadtime demand exceeds the reorder point, a stockout occurs, as illustrated in the following example.

EXAMPLE 7.1 (REVISITED)

Recall that the average leadtime demand is determined to be 20,000 units (see Example 7.1) and the reorder point was set at 20,000 units. Suppose, however, that the manager observes that actual leadtime demand fluctuates between 15,000 and 25,000 units. Because leadtime demand is uncertain, actual leadtime demand is less than 20,000 in some replenishment cycles and larger in others. When the latter situation occurs, we have a stockout.

Let the average leadtime demand be denoted by *LTD* and its standard deviation by σ_{LTD}. Suppose that the reorder point is set at the average leadtime demand, or $ROP = LTD$. Assume further that the distribution (see Appendix II) of leadtime demand is symmetric around its mean. This means that if we carry just enough inventory to satisfy forecast demand during leadtime (with a mean *LTD*), then actual leadtime demand will exceed forecast demand in 50% of our order cycles. We will suffer stockouts, and our service level *SL* will be 50%. Notice that the time interval of interest to measure service level (see section 7.2.1) is the time between the placement of a replenishment order and receipt of that order, which is the lead time. Therefore, *SL* is measured as the probability that the leadtime demand is no greater than the reorder point. To reduce our stockout risk, we may decide to order earlier by setting the reorder point larger than the average leadtime demand. The additional amount that we carry in excess of the average requirements is the safety inventory, denoted by I_{safety}. That is,

$$I_{safety} = ROP - LTD$$

Thus, we have reorder point level expressed as follows:

$$ROP = \text{Average leadtime demand} + \text{safety stock} = LTD + I_{safety} \quad \textbf{(Equation 7.1)}$$

Figure 7.1 illustrates a continuous review reorder point system when the leadtime demand is uncertain. As shown, inventory level fluctuates over time and is not depleted uniformly. Specifically, the on-hand inventory when an order arrives varies between cycles. When actual leadtime demand is smaller than its average value of *LTD* (as in the first cycle in Figure 7.1), the on-hand inventory just before the next order

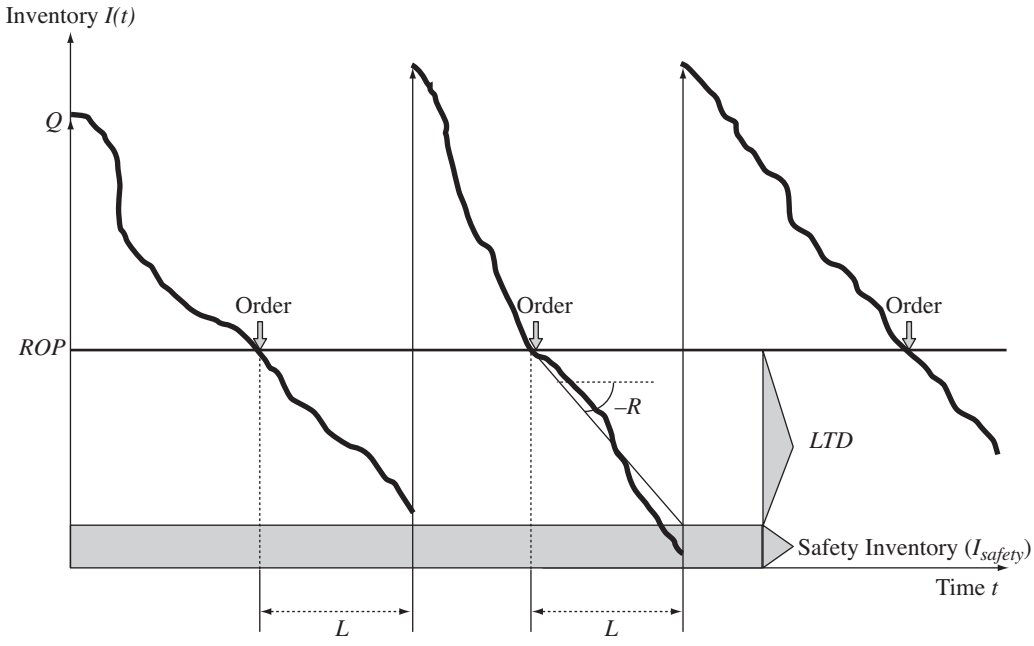


FIGURE 7.1 Continuous Review Reorder Point System

arrives is greater than the safety inventory (I_{safety}). If, however, the actual leadtime demand is larger than its average value of *LTD* (as in the second cycle in Figure 7.1), the on-hand inventory just before the next order arrives is smaller than the safety inventory. Because the average leadtime demand is *LTD*, the average on-hand inventory just before the next order arrives will be equal to the safety inventory (I_{safety}).

Recall from Chapter 6 that average inventory with an order of size Q equals $Q/2$ and is called cycle inventory, I_{cycle}. When leadtime demand is uncertain, we carry safety inventory I_{safety} as well, so that the total average inventory is now

$$I = I_{cycle} + I_{safety} = Q/2 + I_{safety} \qquad \textbf{(Equation 7.2)}$$

Because the average flow rate is R, the average flow time is expressed by Little's law as follows:

$$T = I/R = (Q/2 + I_{safety})/R$$

It represents the average amount of time a typical flow unit waits in inventory before being used. Thus to improve service level by reducing stockout risk calls for an appropriate level of safety inventory, increasing total average inventory and flow time.

7.2.3 Service Level Given Safety Inventory

Service level is measured by the probability (or the proportion of time) that the actual leadtime demand will not exceed the reorder point. Figure 7.2 illustrates the relationship between the distribution of leadtime demand **LTD**, the reorder point *ROP*, and the corresponding service level *SL*. In Figure 7.2, the area under the density curve to the left of the reorder point is the probability *SL* that leadtime demand will be less than the reorder point.

Formally, this area can be written as

$$SL = \text{Prob}(\textbf{LTD} \leq ROP) \qquad \textbf{(Equation 7.3)}$$

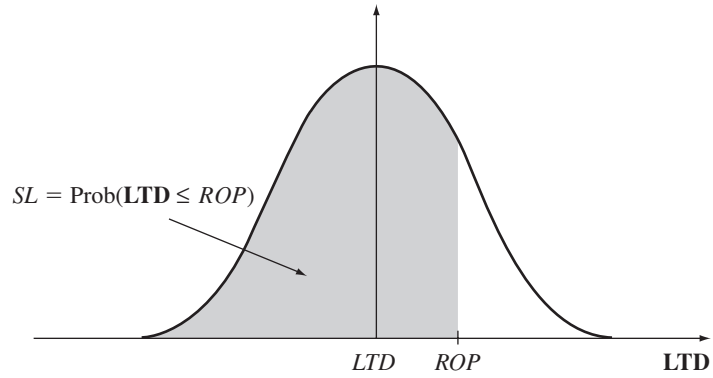


FIGURE 7.2 Reorder Point and Service Level

To compute this probability, we need to know the probability distribution of the random variable **LTD**. It is common to assume that **LTD** is normally distributed with mean *LTD* and standard deviation σ_{LTD}. Thus, the probability density function of **LTD** (representing the probability of different values of **LTD**; see Appendix II for details) is bell shaped—symmetric around *LTD* with a spread representing the magnitude of σ_{LTD}—where larger values of σ_{LTD} correspond to a more dispersed distribution. It can then be shown (see the Appendix at the end of this chapter) that the area covered to the left of the reorder point in the density function for leadtime demand is the same as the area covered to the left of a corresponding constant, represented by z, to the left of a normal density with mean of 0 and standard deviation of 1. Formally,

$$SL = \text{Prob}(\mathbf{LTD} \leq ROP) = \text{Prob}(\mathbf{Z} \leq z)$$

where **Z** is a standard normal random variable with mean 0 and standard deviation 1 and z measures the safety inventory I_{safety} relative to the standard deviation of leadtime demand σ_{LTD}. That is,

$$I_{safety} = z \times \sigma_{LTD} \qquad \textbf{(Equation 7.4)}$$

or

$$z = \frac{I_{safety}}{\sigma_{LTD}}$$

Thus, for any given value of z, the service level *SL* can now be read from the standard normal table given in Appendix II. The service level can also be computed directly in Microsoft Excel as follows:

$$SL = \text{NORMDIST}\,(ROP, LTD, \sigma_{LTD}, \text{True})$$

Example 7.3 illustrates the computation of service level for a given safety inventory.

EXAMPLE 7.3

In Example 7.1, the average leadtime demand for lamps at GE Lighting's Paris warehouse was determined to be 20,000 units. Actual demand, however, varies daily. Suppose, then, that the standard deviation of leadtime demand is estimated to be 5,000 units. The warehouse currently orders a 14-day supply of lamps each time the inventory level drops to 24,000 units. How do we determine service level in terms of the proportion of order cycles over which the warehouse will have stock to meet customer demand? What are the average total inventory and the average flow time?

We know the following: The leadtime demand has mean $LTD = 20{,}000$ units and standard deviation $\sigma_{LTD} = 5{,}000$. Safety inventory can be expressed as follows:

$$\begin{aligned} I_{safety} &= ROP - LTD \\ &= 24{,}000 - 20{,}000 = 4{,}000 \end{aligned}$$

which, when measured as the number of standard deviations, corresponds to

$$z = \frac{I_{safety}}{\sigma_{LTD}}$$

or

$$z = 4{,}000/5{,}000 = 0.8$$

Using the standard normal tables (Appendix II), we now find the service level to be

$$SL = \text{Prob}(\mathbf{Z} \leq 0.8) = 0.7881$$

Alternately, using Microsoft Excel,

$$SL = \text{NORMDIST}\,(24{,}000\,,\, 20{,}000\,,\, 5{,}000\,,\, \text{True}) = 0.7881$$

To summarize, in 78.81% of the order cycles, the warehouse will not have a stockout; alternately, the in-stock probability is 78.81%.

Recall from Example 7.1 that the warehouse manager orders $Q = 28{,}000$ units. Thus, the corresponding cycle inventory

$$I_{cycle} = 28{,}000/2 = 14{,}000$$

Combined with safety inventory

$$I_{safety} = 4{,}000$$

the average total inventory is

$$I = I_{cycle} + I_{safety} = 18{,}000 \text{ units}$$

for an average annual holding cost of

$$€20 \times 18{,}000 = €360{,}000/\text{year}$$

Average flow time, therefore, is

$$T = I/R = 18{,}000/2{,}000 = 9 \text{ days}$$

7.2.4 Safety Inventory Given Service Level

Managers often want to determine the safety inventory and reorder point required to provide a desired level of service. In this case, in Figure 7.2 we know the service level *SL* and want to compute the reorder point *ROP*. To proceed, we must reverse the computational procedure in Section 7.2.3. Knowing *SL*, we first determine the z value from the standard normal tables (Appendix II) such that

$$SL = \text{Prob}(\mathbf{Z} \leq z)$$

Given *SL*, the z value can also be computed directly in Microsoft Excel as follows:

$$z = \text{NORMSINV}\,(SL)$$

We can then compute the safety inventory

$$I_{safety} = z \times \sigma_{LTD}$$

and then the reorder point

$$ROP = LTD + I_{safety}$$

Alternately, the reorder point can be directly computed using Microsoft Excel as

$$ROP = \text{NORMINV}(SL, LTD, \sigma_{LTD})$$

Thus, to determine the reorder point for a desired service level, we need information regarding average leadtime demand *LTD* and its standard deviation σ_{LTD}. These figures in turn will depend on flow rate **R** (its average and standard deviation) and leadtime of supply **L** (its average and standard deviation). To keep our focus on the interaction between service levels and safety inventory, we assume in this section that the average and the standard deviation of leadtime demand are known. We discuss methods for estimating information about leadtime demands in Section 7.5. Example 7.4 illustrates the computation of the safety inventory and reorder point to achieve a given service level.

EXAMPLE 7.4

Reconsider Example 7.3. We determined that with a safety inventory of 4,000 units, the provided service level was 78.81%. Recently, customers of the Paris warehouse have started complaining about the frequent stockout of lamps when they placed their orders with the warehouse. In response, the warehouse manager is considering increasing the service level but does not know how much the increase may cost in extra inventory. He wants to evaluate the cost of providing service levels of 85%, 90%, 95%, and 99%. How will he determine how much safety inventory should be carried to provide these levels?

Recall first that the average (*LTD*) and standard deviation (σ_{LTD}) of the leadtime demand were 20,000 and 5,000 units, respectively. Now consider a service level of 85%. To determine the corresponding value of *z*, we must find that value of *z* such that

$$\text{Prob}(\mathbf{Z} \leq z) = 0.85$$

Using the standard normal tables, one can read the *z* value for 85% service level as 1.04. Alternately, in Microsoft Excel, we will write

$$z = \text{NORMSINV}(0.85) = 1.04$$

Safety inventory is therefore $I_{safety} = z \times \sigma_{LTD} = 1.04 \times 5{,}000 = 5{,}200$ units, and the reorder point is $ROP = LTD + I_{safety} = 20{,}000 + 5{,}200 = 25{,}200$ units. We repeat this process for each desired service level—reading the *z* value, computing the safety inventory, and calculating the reorder point. The results are summarized in Table 7.2, where we observe that required safety inventory increases with service level. Whereas an increase of 5% in service, from 85% to 90%, requires an additional safety inventory of 1,206 units, the next 5% increase in service level, from 90% to 95%, requires an additional safety inventory of 1,840 units. Thus we observe a nonlinear relationship between safety inventory and service level.

Table 7.2 Safety Inventory versus Service Level

Service Level (*SL*)	*z* Value	Safety Inventory (I_{safety})	Reorder Point (*ROP*)
85%	1.04	5,200	25,200
90%	1.28	6,406	26,406
95%	1.65	8,246	28,246
99%	2.33	11,686	31,686

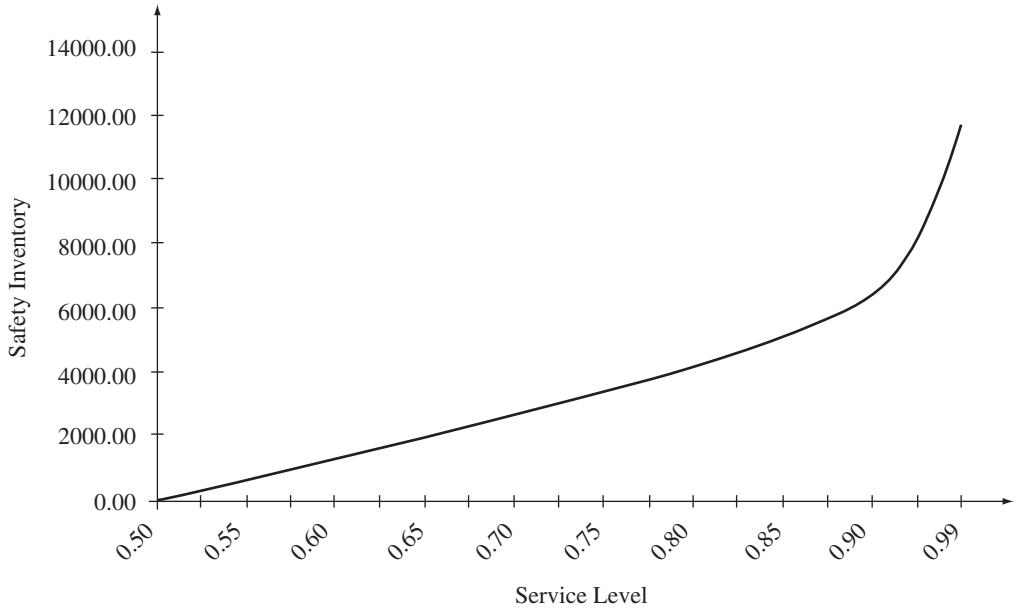


FIGURE 7.3 Safety Inventory versus Service Level

Increasing service level increases the required safety inventory *more than proportionately*—as seen in Figure 7.3. Because providing higher levels of service gets increasingly more expensive in terms of inventory holding cost, the process of selecting service and safety inventory levels is an important strategic decision. A firm may choose, for instance, to provide high-quality service in terms of product availability, or it may choose to be a low-cost provider by holding down inventory costs. In either case, it is positioning itself along the service versus cost trade-off curve displayed in Figure 7.3. Our aim in providing Example 7.4 and Table 7.2 is to show how that position could be operationalized. For example, if the warehouse decides to position itself as a company providing high service at a 99% level, then it must carry a safety inventory of approximately 11,686 units along with the entailing cost.

7.3 OPTIMAL SERVICE LEVEL: THE NEWSVENDOR PROBLEM

Thus far, we have derived safety inventory for a desired level of customer service. However, what level of service should a firm offer? An optimal service level should balance the benefits of improved service in terms of supply continuity and customer satisfaction with the additional costs of holding required safety inventory. In this section, we analyze this problem in a simpler context with a problem involving only one order cycle. The qualitative principles that emerge in the upcoming discussion carry over to this discussion of the reorder point model.

A large proportion of retail sales involves "fashion goods" with short product life cycles of a few months. Usually, the retailer has only one or two buying opportunities, and at the end of the product life cycle, remaining items must be marked down for sale or even disposed of at a loss. Newspapers and magazines, for example, have limited lives (a day, a week, a month) at the end of which they lose most of their value. Perishable grocery items—fish, produce, bread, and milk—also have limited shelf lives and must be discarded after expiration dates. Seasonal items like Christmas trees, snow blowers and lawn mowers, and summer and winter apparel are bought and sold only at certain times of the year. In these cases, purchasing too few or too many items in relation

to uncertain customer demand entails tangible costs. Because margins are usually high before the end of the season, retailers with insufficient inventory lose potential profits. Conversely, because postseason markdowns can be significant, those with excess inventory lose money through lower margins.

Thus, it is important to perform an economic analysis in order to determine the optimal order quantity. Such an analysis should balance the expected costs of ordering too little (such as customer dissatisfaction and the opportunity cost of lost revenue) with the expected costs of ordering too much (such as markdowns and disposal costs). In the operations literature, this *basic model of decision making under uncertainty whereby the decision maker balances the expected costs of ordering too much with the expected costs of ordering too little to determine the optimal order quantity* is discussed as the **newsvendor problem**. It differs from the EOQ inventory model (illustrated in Chapter 6), which focuses on scale economies and, more importantly, assumes no uncertainty. The newsvendor model, which is a basic model for decision making under uncertainty, highlights the role of uncertainty, assumes no scale economies, and boasts a wide variety of applications, as illustrated at the end of this section. Let us consider Example 7.5.

EXAMPLE 7.5

Big George Appliances[2] is an electronics superstore in Ann Arbor, Michigan. It sells consumer electronics items as well as appliances. Big George is considering carrying a 54" plasma HDTV for the upcoming Christmas holiday sales. Each HDTV can be sold at $2,500. Big George can purchase each unit for $1,800. Any unsold TVs can be salvaged, through end-of-year sales, for $1,700. The retailer estimates that the demand for this new HDTV will be between 100 and 200 units with probability weights as given in Table 7.3. Big George needs to determine the number of HDTVs to be purchased for this season's sales.

We use the following notation in Table 7.3. The uncertain demand for HDTV is represented by **R**. The variable **R** can take various values, denoted by *r,* ranging from 100 to 200; this is column 1. The probability of a particular demand value *r* is given by

Table 7.3 Demand for HDTV at Big George

Demand r	Probability Prob ($R = r$)	Cumulative Probability Prob ($R \leq r$)	Complementary Cumulative Probability Prob ($R > r$)
100	0.02	0.02	0.98
110	0.05	0.07	0.93
120	0.08	0.15	0.85
130	0.09	0.24	0.76
140	0.11	0.35	0.65
150	0.16	0.51	0.49
160	0.20	0.71	0.29
170	0.15	0.86	0.14
180	0.08	0.94	0.06
190	0.05	0.99	0.01
200	0.01	1	0

[2]All numbers in the example are fictitious and used only to illustrate the concepts.

Prob($\mathbf{R} = r$) and is listed in column 2. The cumulative probability, written as Prob($\mathbf{R} \leq r$), representing the chance that demand **R** will be less than or equal to a particular value *r*, is given in column 3. Finally, the last column gives the complementary cumulative probability, written as Prob($\mathbf{R} > r$), which is the probability that the demand **R** will exceed a particular value *r*. Using data from Table 7.3, we can compute the average demand as the weighted average of all possible demand values between 100 and 200 and their respective probabilities. Let *R* represent this average. Notationally, if *x* takes values from 1 to *k*, we write $\sum_{x-1}^{k} x$ to represent the sum of all values of *x* ranging from 1 to *k*. Using this notation, we write

$$R = \sum_{demand=100}^{200} demand \times probability$$

$$R = \sum_{r=100}^{200} r \times f(r) = 151.60 \text{ units}$$

This estimate of the average demand, *R*, represents the forecast of sales. If there were no uncertainty in demand for the HDTVs, then Big George should purchase 152 units. With uncertain demand, however, there is a 49% probability that actual demand will exceed 150, resulting in a stockout and lost revenue. There is also a 51% chance that at least one HDTV will be left over to be salvaged at a loss. Thus, ordering the mean demand may not maximize profitability.

To facilitate the determination of the optimal order quantity, we first outline a procedure to estimate the expected profits for a particular order quantity, say, $Q = 160$ units. First, we recall the following facts:

- If actual demand is 160 or higher, all 160 units will be sold at a profit of \$700 each.
- If the demand is fewer than 160 units, some of the 160 units will have to be disposed of at a loss of \$100 each (the difference between the purchase price and the salvage value).

Thus, every unit sold fetches a unit profit of \$700, and every unsold unit costs \$100. Given an order quantity, we can compute the gross profit for every possible demand scenario (ranging from demand values of 100 to 200). Each demand scenario, however, occurs with a known probability, as given in Table 7.3. For example, if the order quantity is 160 and the demand realized is 100, Big George will then sell 100 units at a profit of \$700 each. However, 60 units will be left unsold, incurring a loss of \$100 each. The gross profit for demand value of 100 is then ($100 \times \$700 - 60 \times \$100$) = \$64,000. The chance of a demand realization of 100 is Prob($\mathbf{R} = 100$) = 0.02. Observe that under each of the scenarios, when demand realized is greater than or equal to 160 units, sales will be 160 units, and there will be no excess units left over, giving gross profits of $160 \times \$700 = \$112{,}000$. These scenarios, cumulatively, will occur with probability Prob($\mathbf{R} > 160$). By multiplying the gross profit under a given scenario with its probability and then summing across all possible scenarios, we can compute the expected profit of a given order quantity. For an order quantity of 160 units, the expected profit is computed as follows:

$$\begin{aligned}(100 \times 700 - 60 \times 100)\ \text{Prob}(\mathbf{R} = 100) + (110 \times 700 - 50 \times 100)\ \text{Prob}(\mathbf{R} = 110)\\ + (120 \times 700 - 40 \times 100)\ \text{Prob}(\mathbf{R} = 120) + \ldots + (160 \times 700)\ \text{Prob}(\mathbf{R} \geq 160)\\ = \$101{,}280\end{aligned}$$

A similar approach can be used to determine the expected profit resulting from an order quantity of $Q = 110, \ldots, 200$. The expected profits for various order quantities are displayed in Table 7.4.

Table 7.4 Order Quantity versus Expected Profits

Order Quantity (Q)	Expected Profit
100	$70,000
110	$76,840
120	$83,280
130	$89,080
140	$94,160
150	$98,360
160	$101,280
170	$102,600
180	$102,720
190	$102,200
200	$101,280

The order quantity that yields maximum profit equals 180 units—which is our desired order quantity. The optimal order size is larger than expected demand because with uncertain demand, we do not simply order the expected value of demand. Rather, our decision depends on a broader range of economic considerations, including price, purchasing cost, and salvage value of the unit.

The generic problem can be stated as follows: Consider a retailer who sells ski parkas. Let **R** denote the uncertain demand for this product. Every ski parka sold during the season fetches retail price of p per unit. Any parka not sold during the season can be disposed of at a markdown price of v per unit. The unit purchase cost (wholesale price paid by the retailer) of one parka is c. The retailer must decide how many parkas to order. Suppose the retailer decides to order Q parkas. As a consequence, the various cash flows can be described as follows:

- ***In-season sales:*** The number of parkas sold during the season will depend on the realized demand and will be given by the lesser of the demand and the quantity stocked and is equal to min (Q, **R**). Each of these parkas generate a revenue of p, giving a total revenue $p \times min$ (Q, **R**).
- ***Markdown sales:*** Parkas not sold during the regular season will be salvaged at the end of the season. The number of parkas left over at the end of the season is given by max ($Q - \mathbf{R}$, 0), each earning a revenue of v. The total revenues earned from markdown sales is then $v \times max$ ($Q - \mathbf{R}$, 0).
- ***Purchase cost:*** Finally, the retailer purchases Q parkas at a unit cost of c per unit, resulting in a total purchase cost of cQ.

The realized value of in-season sales and markdown sales will differ depending on the demand that materializes. The retailer, however, has to make her decision before observing the demand. She chooses an order quantity Q that optimizes the expected value of the profits given as

$$\text{Expected profit} = \text{Expected in-season sales} + \text{Expected markdown sales} - \text{Purchase cost}$$

Marginal Analysis A more insightful approach to understanding the trade-offs involved in deciding optimal order quantity entails **marginal analysis**: *comparing expected costs and benefits of purchasing each incremental unit.* First we must define the following:

- **Net marginal benefit** is *the difference between the unit price of the product and unit marginal cost of procurement.* The net marginal benefit from each additional unit,

denoted by *MB*, is its contribution margin. If the unit retail price is p and the unit purchase cost is c, then

$$MB = p - c$$

In practice, it may also include the opportunity cost of lost goodwill had the unit not been stocked but were demanded.

- **Net marginal cost** is *the difference between unit marginal cost of procurement and its salvage value.* The net marginal cost of stocking an additional unit, denoted by *MC*, is the effective cost if the unit remains unsold under conditions of low demand. If the unit salvage value is v and the purchase cost is c, then

$$MC = c - v$$

We receive the net marginal benefit only when the additional unit sells, which will occur whenever demand exceeds Q. At any order quantity, Q, the expected marginal benefit from ordering an additional unit is

$$MB \times \text{Prob}(\mathbf{R} > Q)$$

At the same time, we suffer the net marginal cost only when the additional unit does not sell, which will occur whenever demand is no more than Q. The expected marginal cost of having a unit left over is

$$MC \times \text{Prob}(\mathbf{R} \leq Q)$$

Note that while the expected marginal benefit from purchasing an additional unit is decreasing, expected marginal cost is increasing in the order quantity Q. As long as expected benefit is greater than expected cost, Q should be increased until the reverse is true. Thus, the optimal Q is the first value Q^* for which the expected cost of ordering an additional unit exceeds the expected benefit; that is,

$$MC \times \text{Prob}(\mathbf{R} \leq Q^*) \geq MB \times \text{Prob}(\mathbf{R} > Q^*)$$

Since

$$\text{Prob}(\mathbf{R} > Q) = 1 - \text{Prob}(\mathbf{R} \leq Q)$$

the condition for optimality of Q^* can be rewritten as follows:

$$MC \times \text{Prob}(\mathbf{R} \leq Q^*) \geq MB \times [1 - \text{Prob}(\mathbf{R} \leq Q^*)]$$

Rearranging terms, we arrive at an optimal order quantity as *the smallest value* Q^* *such that*

$$\text{Prob}(\mathbf{R}) \leq Q^* \geq \frac{MB}{(MB + MC)}$$

Thus, computing optimal order quantity is a two-step procedure:

1. Compute the ratio $\dfrac{MB}{(MB + MC)}$.
2. Determine optimal order quantity, Q^*, from the cumulative distribution of demand $\mathbf{R}$.

We illustrate this procedure in Example 7.6.

EXAMPLE 7.6

We now apply these principles to the problem of ordering HDTVs for Big George Appliances. Recall that

$$MB = p - c = 2{,}500 - 1{,}800 = \$700$$

and

$$MC = c - v = 1{,}800 - 1{,}700 = \$100$$

Thus,

$$\frac{MB}{MB + MC} = \frac{700}{700 + 100} = 0.875$$

From the cumulative distribution of demand in Table 7.3 (column 3), we find that

$$\text{Prob}(\mathbf{R} \leq 180) = 0.94$$

and

$$\text{Prob}(\mathbf{R} \leq 170) = 0.86$$

Therefore, the smallest Q^* such that $\text{Prob}(\mathbf{R} \leq Q^*) \geq 0.875$ is $Q^* = 180$.

We can simplify this procedure even further. It is often more convenient, for instance, to assume that demand is a continuous random variable, whereby all (non-integer) values of **R** and Q become possible. If we make this assumption, then at optimal Q we can exactly balance out the marginal benefit of increasing Q (by a fractional amount) with the loss of keeping Q at its current level. Thus,

$$MC \times \text{Prob}(\mathbf{R} \leq \text{Q*}) = MB \times [1 - \text{Prob}(\mathbf{R} \leq \text{Q*})]$$

which gives us an optimal order quantity, Q^*, that satisfies the following relationship:

$$\text{Prob}(\mathbf{R} \leq Q^*) = \frac{MB}{(MB + MC)}$$

Recall from Section 7.2 that cycle service level was defined as the probability of not stocking out in a cycle. If demand is represented by **R** and order quantity by Q, then cycle service level is $\text{Prob}(\mathbf{R} \leq Q)$. Because Q^* is optimal order quantity determined by the economic trade-off between costs of under- and overstocking, $\text{Prob}(\mathbf{R} \leq Q^*)$ is the optimal probability of not stocking out. Therefore, using the earlier relationship for Q^*, the optimal service level SL^* is given by the following formula:

$$\textit{Newsvendor formula: } SL^* = \text{Prob}(\mathbf{R} \leq Q^*) = \frac{MB}{MB + MC} \quad \textbf{(Equation 7.5)}$$

Note that optimal service level depends only on the net marginal benefit and cost of stocking a unit and not on the probability distribution function. Furthermore, it increases with the net marginal benefit, MB, and decreases with the net marginal cost, MC. Thus, the more expensive the stockouts and/or the lower the cost of disposing of excessive inventory, the higher the optimal service level. For Examples 7.5 and 7.6, optimal service level is computed as equal to

$$\frac{MB}{MB + MC} = 0.875$$

Knowing our optimal service level, we can now determine optimal order quantity from the probability distribution of demand. Let us assume, for example, that demand is normally distributed with mean R and standard deviation σ_R. In that case, the optimal order quantity Q^* can be determined in one of two ways:

1. From the standard normal tables given in the Appendix II, we first determine z corresponding to the optimal service level SL^* and then compute

$$Q^* = R + z \times \sigma_R$$

2. Or, from the Microsoft Excel function,

$$Q^* = \text{NORMINV}\,(SL^*, R, \sigma_R)$$

We illustrate this computation in Example 7.7.

EXAMPLE 7.7

Recall that in Example 7.5, $R = 151.60$. The variance of R can be computed as the average squared deviation from its mean, or

$$\sigma_R^2 = \sum_{r=100}^{200} [(r - R)^2 \times \text{Prob}(\mathbf{R} = r)] = 503.44$$

Taking its square root gives the standard deviation $\sigma_R = 22.44$ units—a figure that measures the variation in actual demand around its mean. We need Q^* such that

$$\text{Prob}(\mathbf{R} \leq Q^*) = 0.875$$

Looking up the normal tables, we find that $z = 1.15$ and

$$Q^* = R + z \times \sigma_R = 151.60 + 1.15 \times 22.44 = 177.41$$

Alternately, using the Microsoft Excel formula, we get

$$Q^* = \text{NORMINV}\,(0.875\,,\,151.60\,,\,22.44) = 177.41$$

which is close to our earlier answer of 180 units. The discrepancy arises because we approximated discrete demand probability density in Example 7.5 with a continuous probability density.

In the newsvendor model, the difference between the optimal order quantity Q^* and the mean demand R is the single-order equivalent of safety inventory I_{safety} that we considered in the preceding section. Thus, the qualitative conclusions of this section apply to the preceding discussion. Of course, the economics of the situation could be such that the optimal order quantity is below the mean demand, in which case we say that the firm carries a negative safety stock. This will occur, for example, when optimal service level is below 50 percent. With uncertain demand, therefore, we determine optimal service level—and corresponding safety inventory—by balancing the expected marginal benefit of an additional unit with those of expected marginal cost. Intuitively, we rationalize that if the net marginal benefit is twice the net marginal cost, we need an order quantity that gives us a probability of overstocking that is twice the probability of understocking. To summarize, the optimal service level increases with the net marginal benefit and decreases with the net marginal cost. The order quantity increases with the optimal service level and the mean and the standard deviation of demand.

The newsvendor model is a fundamental model for decision making under uncertainty and has applications in a wide variety of areas. Consider the following:

1. AT&T offers two types of data plans for laptops. DataConnect 200MB costs $35 per month for data usage upto 200MB, with extra charge of $0.10 per additional MB for data usage within the United States. Similarly, DataConnect 5GB costs $60 per month for data usage upto 5GB, with an extra charge of $0.05 per additional MB of data usage. Which plan should you sign up for?
2. *Avatar* has just been released on Blu-Ray for the rental market. Netflix needs to place orders for the disc. The studio charges a unit price for each disc Netflix purchases. The rental lifetime of a typical tape is about four weeks. Netflix can rent

the tape several times during this period and then sell it off at a steep discount as a pre-viewed disc. How many Blu-Ray discs of *Avatar* should Netflix stock?

3. The IRS code allows employees to set aside a certain part of their salary into a health care Flexible Savings Account (FSA) earmarked for health care expenses. The amount set aside for a given year must be committed to by the end of the previous year. During the year, an employee can use these moneys, tax free, to cover qualified health care expenses. However, any amount not claimed during the year is forfeited. The health care needs of a family cannot be accurately forecasted. How much money should an employee set aside in their FSA?
4. Delta needs to determine the number of reservations to accept for the 7:00 a.m. flight from Detroit to San Francisco. The plane has a capacity of 350 seats. Passengers always make last-minute changes to their travel plans, resulting in cancellation of their reservations. Therefore, accepting exactly 350 reservations may result in some unused seats because of cancellations. If more than 350 reservations are taken and everyone shows up, then some passengers need to be bumped at a cost. How many reservations should Delta take for this flight?
5. Amgen is gearing up for the introduction of Enbrel, a breakthrough drug for arthritis. There is a wide range of estimates for the potential market for Enbrel. However, because of the long lead time for building a plant, Amgen must decide on capacity long before the product is launched. If demand ultimately outstrips capacity, there is the potential of lost revenue or of excessive costs of subcontracting. On the other hand, too little demand will result in capital tied up in unused capacity. How much capacity should Amgen build?

7.4 LEADTIME DEMAND VARIABILITY

The rest of this chapter considers sources of demand variability and operational strategies for reducing this variability.

Recall that leadtime demand **LTD** refers to the flow unit requirement from the time an order is placed until it is received. We carry safety inventory to satisfy this requirement a proportion of time corresponding to the service level. As discussed, both the safety inventory and the service level depend critically on the variability in the leadtime demand—if the leadtime demand were constant and known, we could guarantee 100 percent service level with no safety inventory. In this section, we consider factors that affect the service level and the safety inventory by contributing to variability in the leadtime demand.

7.4.1 Fixed Replenishment Lead Time

For the sake of simplicity, we first consider the case of known (fixed) replenishment lead time L measured in periods (days, weeks, months). We assume that our supplier is perfectly reliable, and we postpone the discussion of variability in leadtimes to Section 7.4.2.

First, let $\mathbf{R}_t$ denote (uncertain) demand in period t. For a supply lead time of L number of periods, total leadtime demand will be

$$\mathbf{LTD} = \mathbf{R}_1 + \mathbf{R}_2 + \ldots + \mathbf{R}_L$$

We will assume that demand levels between periods are independent and follow the same distribution—that is, they are independent and identically distributed random variables. Average leadtime demand, therefore, will be given by

$$LTD = L \times R \qquad \textbf{(Equation 7.6)}$$

where L is lead time in number of periods and R is average demand per period. Since L is constant, variability in the leadtime demand arises from variability in the periodic

demand. As noted, the statistical measure of variability is its standard deviation. Let σ_R be the standard deviation of demand (flow rate) per period (day, week, or month). Statistically, to compute the standard deviation σ_{LTD} of the leadtime demand, it is convenient to first estimate the variance of the leadtime demand given by σ^2_{LTD} as

$$\sigma^2_{LTD} = \sigma^2_R + \sigma^2_R + \cdots + \sigma^2_R = L \times \sigma^2_R$$

This follows from the fact that the variance of the sum of L independent random variables equals the sum of their variances. Thus, standard deviation of leadtime demand is

$$\sigma_{LTD} = \sqrt{L} \times \sigma_R \qquad \textbf{(Equation 7.7)}$$

If we know the leadtime of supply and the variability in demand per period, we can compute safety inventory to achieve a desired level of service using ideas discussed in Section 7.2.4.

In addition to its dependence on service level (discussed in Section 7.2), safety inventory also depends on the standard deviation of leadtime demand, which depends in turn on both length of supply leadtime and variability in demand. Specifically, greater variability in leadtime demand results from longer leadtime, more variable demand per period, or both. More safety inventory is also needed to provide a desired level of service.

EXAMPLE 7.8

GE Lighting's Paris warehouse manager wants to know if he can reduce procurement costs. The transportation department has proposed that material be shipped by ocean freight, which will reduce the per unit cost but increase the replenishment lead time to 20 days from the present 10 days. The manager needs to know the ramifications of this proposal. What impact, if any, would the new proposal have on the inventory carried in the warehouse?

We proceed in the following manner. Recall from Example 7.1 that the average daily demand R is 2,000 units. Recall also from Example 7.3 that the standard deviation of leadtime demand was specified as 5000. That is, we have $\sigma_{LTD} = 5000$. Then using Equation 7.7 with a replenishment lead time, $L = 10$ days, we estimate $\sigma_R = 1{,}581$ as the standard deviation of daily demand. For the new leadtime of $L = 20$ days, we can compute the standard deviation of the leadtime demand as follows:

$$\sigma_{LTD} = \sqrt{L} \times \sigma_R = \sqrt{20} \times 1{,}581 = 7{,}070$$

For a 95% service level, required safety inventory is expressed as

$$I_{safety} = z \times \sigma_{LTD} = 1.65 \times 7{,}070 = 11{,}666 \text{ units}$$

For a similar service level, when replenishment lead time was 10 days, the safety inventory was estimated in Example 7.4 to be 8,246 units. Thus, under the new proposal, the safety inventory increases by 3,420 units (or 41.4%) from 8,246 to 11,666 because of an increase in replenishment lead time. The additional cost of this inventory has to be traded off with any reduction in transportation cost to determine whether to accept the new proposal.

Example 7.8 illustrates the connection between replenishment lead time and safety inventory. This has managerial implications covering several issues, such as transportation mode choice and supplier selection and sourcing. A decision on transportation mode choice, such as air freight versus ocean freight, impacts the replenishment lead time. This in turn affects the safety inventory necessary to provide a specific service level, as shown in Example 7.8.

Similarly, from a sourcing perspective, suppose we can choose between two suppliers, one that offers a lower price but a longer lead time and the other that offers a shorter time but a higher price. In such situations, selecting a supplier on the basis of the price alone could result in the need to carry larger safety inventory—a decision that may increase total cost. Both contexts, then, call for a decision-making framework based on total costs, including material, transportation, and inventory costs.

7.4.2 Variability in Replenishment Lead Time

In addition to its duration, variability in lead time is also an important contributor to variability in the leadtime demand. To develop some intuition for the effect of variability in lead time, suppose that while demand rate R is fixed and known, lead time is a random variable, **L**, with mean L and standard deviation σ_L. In this case, uncertain leadtime demand is expressed as

$$\mathbf{LTD} = R \times \mathbf{L}$$

It has mean

$$LTD = R \times L$$

and variance

$$\sigma_{LTD}^2 = R^2 \times \sigma_L^2$$

The last expression follows from the fact that the variance of a constant multiplied by a random variable is equal to the square of that constant times the variance of the random variable.

More generally, suppose that both demand rate **R** and lead time **L** are random variables. If so, the leadtime demand is the sum of a random number of random variables. To compute the required safety inventory, therefore, we must compute the mean and the variance of the leadtime demand. Since the average leadtime is L and average flow rate is R, it is clear that the average leadtime demand is

$$LTD = L \times R$$

Assuming that the demand and replenishment lead times are independent random variables, the variance of the leadtime demand can be computed by combining two special cases:

1. Variance of the leadtime demand when flow rate is random but the lead time is fixed (a situation discussed in the previous section)
2. Variance of the leadtime demand when the flow rate is constant but lead time is random (a situation discussed at the beginning of this section)

Total variability in the leadtime demand is then the sum of the two individual effects:

$$\sigma_{LTD}^2 = L\sigma_R^2 + R^2\sigma_L^2$$

The standard deviation of the leadtime demand is then computed by taking the square root of the variance of the leadtime demand and is given by the following formula:

$$\sigma_{LTD} = \sqrt{L\sigma_R^2 + R^2\sigma_L^2} \qquad \textbf{(Equation 7.8)}$$

The exact derivation of this intuitive explanation can be found in Ross (1972). The impact of variability in lead time on safety inventory is illustrated in Example 7.9.

EXAMPLE 7.9

Return for a moment to Example 7.1, and suppose that the replenishment lead time has recently become more variable. Specifically, suppose that the replenishment lead time has a mean of 10 days and a standard deviation of 2 days (with all remaining data as specified in Example 7.1). How much safety inventory does the Paris warehouse need in order to provide a 95% service level?

Again, we start with the following data:

$$L = 10\,,\ \sigma_L = 2\,,\ R = 2{,}000\,,\ \sigma_R = 1{,}581$$

Thus, we see that

$$\sigma^2_{LTD} = (10)(1{,}581)^2 + (2{,}000)^2(2)^2 = 40{,}995{,}610$$

Taking the square root, we get

$$\sigma_{LTD} = 6{,}402.78$$

Therefore, safety inventory must be

$$I_{safety} = 1.65 \times \sigma_{LTD} = 10{,}565 \text{ units}\,,$$

a significantly higher number compared with only 8,246 units needed if lead time were *exactly* 10 days.

We can arrive at an intuitive understanding of this increase. With variability in lead time, it is likely that actual lead time of supply will often be larger than 10 days. The process manager must now account for this increase by carrying more safety inventory.

Thus, variability in lead time of supply increases the safety inventory (and flow time). Reliable suppliers who make on-time deliveries contribute directly to a firm's bottom line and level of customer service.

In summary, we have shown how uncertainty in demand and supply affects raw material and product availability. To provide better service in the face of uncertainty, firms carry safety inventory. Three key factors affect the amount of safety inventory that a company carries under given circumstances:

1. Level of customer service desired
2. The average and the uncertainty in demand
3. The average and the uncertainty in replenishment lead time

In turn, there are two primary levers for reducing the level of safety inventory:

1. Reducing both the average and standard deviation of replenishment lead time
2. Reducing demand variability

Although improved forecasting can reduce variability in demand, too many firms tend to think it is their only option. Better forecasting can help, but it is not a panacea. As discussed, reducing the lead time and reducing its variability are also important levers. In Sections 7.5 and 7.6, we explore two further ways of reducing variability: aggregating demand and using shorter-range forecasts.

7.5 POOLING EFFICIENCY THROUGH AGGREGATION

Recall from Section 7.1 the third characteristic of forecasts: Aggregate forecasts are more accurate than individual forecasts. The basic concept of **aggregation**—*pooling demand for several similar products*—can be applied broadly. Indeed, firms often aggregate sales according to various geographical regions and/or types of products. Improved forecast

accuracy due to aggregation is simply a statistical property, and we can devise important operational strategies to exploit this property in effective inventory management.

7.5.1 Physical Centralization

Suppose a firm stocks its product in multiple warehouses to serve geographically dispersed customers. Because all the locations face uncertain demand, each should carry some safety inventory. Assume that the company's warehousing operations are decentralized—that each warehouse operates independently of the others. It is possible, then, that one warehouse will be out of stock while another has the product in stock. Although the total distribution system has sufficient inventory, it may be held at the wrong location. As a result of this imbalance of inventory, some customer demand may not be satisfied.

Suppose, however, *that the firm can consolidate all its stock in one location from which it can serve all its customers.* We will call this alternative system the **physical centralization** of inventory. Because centralization eliminates the possibility of stock imbalance, all customer demand will be met as long as there is inventory in the system. The centralized system, therefore, will provide *better* customer service than the decentralized network and will do so with the *same* total inventory. Equivalently, to provide the same level of service as in the decentralized system, the centralized system would need *less* inventory.

Let us make these claims more precise. Suppose that a firm serves locations 1 and 2, and assume that the respective leadtime demands—$\mathbf{LTD}_1$ and $\mathbf{LTD}_2$—are statistically identically distributed, each with mean of *LTD* and standard deviation of σ_{LTD}. To provide a desired level of service *SL*, each location must carry safety inventory

$$I_{safety} = z \times \sigma_{LTD}$$

where z is determined by the desired service level (as discussed in Section 7.2). If each facility faces identical demand and provides identical service levels, the total safety inventory in the decentralized system is equal to $2z\sigma_{LTD}$.

INDEPENDENT DEMANDS

Consider centralizing the two inventories in one location when leadtime demands at the two locations are independent. This centralized pool will now serve the total leadtime demand

$$\mathbf{LTD} = \mathbf{LTD}_1 + \mathbf{LTD}_2$$

Recall that the mean and the variance of a sum of independent random variables are, respectively, equal to the sum of their means and variances. The mean of total leadtime demand faced by the central warehouse is thus

$$LTD + LTD = 2\,LTD$$

Its variance is

$$\sigma^2_{LTD} + \sigma^2_{LTD} = 2\sigma^2_{LTD}$$

The standard deviation of the leadtime demand at the centralized location, therefore, is σ_{LTD}. Note that although consolidation of demands doubles the mean, the standard deviation increases only by a factor of $\sqrt{2}$ or 1.414.

Intuitively, we understand that high and low demands in the two locations will tend to counterbalance each other, thereby yielding a more stable total demand. Safety inventory carried in the centralized system is then equal to

$$I^c_{safety} = z \times \sqrt{2}\sigma_{LTD}$$

Comparing the safety inventories carried by decentralized (I^d_{safety}) and centralized (I^c_{safety}) systems, we observe that when both systems offer the same level of service, the total safety inventory required by the centralized operation is $\sqrt{2}$ times the required total safety inventory in the decentralized operation. That is, the safety inventory in a centralized system is less than in a two-location decentralized system by a factor of $\frac{1}{\sqrt{2}}$. We can generalize our analysis of the benefits of centralizing two locations to consider the centralization of N locations. The safety inventory needed when N locations are centralized is given by

$$I^c_{safety} = z \times \sqrt{N}\sigma_{LTD} \qquad \textbf{(Equation 7.9)}$$

A similar N location decentralized network will require a safety inventory investment of N times the safety inventory in each warehouse, which will total $N \times z \times \sigma_{LTD}$. Thus, centralization will reduce safety inventory investment by a factor of $\sqrt{N}$. The concept of centralization is illustrated in Example 7.10.

EXAMPLE 7.10

Recall that GE Lighting operated seven warehouses. An internal task force had recommended that it consolidate all the seven warehouses into one central warehouse to serve customers in Austria, Belgium, France, Germany, the Netherlands, Luxembourg, and Switzerland. Assume that the replenishment lead time to this central warehouse will remain at ten days. What will be the impact of accepting the task force recommendations?

In the current decentralized system, each warehouse orders independently of the other warehouses. In Example 7.4, we estimated that a warehouse facing average leadtime demand of 20,000 units with a standard deviation of 5,000 units needs to carry a safety inventory of $I_{safety} = 8{,}246$ to provide a 95% service level. Assuming that each of the other warehouses faces similar demand patterns and wants to offer the same service level, the total safety inventory carried across the seven warehouses will be

$$I^d_{safety} = 7 \times 8246 = 57{,}722.$$

If the task force recommendation is accepted, the single central warehouse will face a total leadtime demand with mean and standard deviation of

$$LTD = 7 \times 20{,}000 = 140{,}000$$

$$\sigma_{LTD} = \sqrt{7} \times 5000 = 13{,}228.8$$

To provide a 95% service level, the central warehouse must carry a safety inventory:

$$I^c_{safety} = z \times \sigma_{LTD} = 1.65 \times 13{,}228.8 = 21{,}828$$

Thus, we see that the required safety inventory with the single central warehouse is 35,894 less than that required under the current decentralized network of seven warehouses. This reduction represents a decrease in safety inventory of 62%, or reduction by a factor of $\sqrt{7}$ or 2.65.

Square Root Law The savings illustrated in Example 7.10 results from the **square root law**, which states that *total safety inventory required to provide a specified level of service increases by the square root of the number of locations in which it is held.* This principle is displayed graphically in Figure 7.4.

In addition to the benefits of reducing the safety inventory, centralization also reduces the cycle inventory, as we saw in Chapter 6. The reduction in cycle inventory

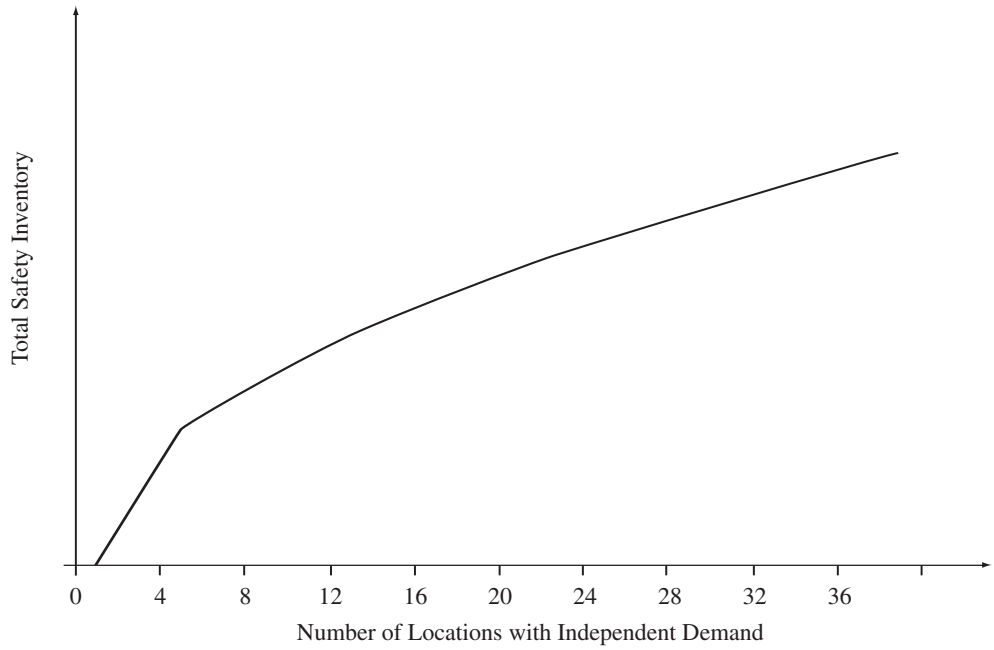


FIGURE 7.4 Square Root Law of Pooling

results from the fact that centralization allows better use of economies of scale in procurement and production. Physical centralization is a common practice for retailers with catalog and mail-, telephone-, or Internet-order operations.

Correlated Demands In the previous discussion, we have shown the benefits of centralization when demands in the various locations were *independent.* Does centralization offer similar benefits when demands in the multiple locations are correlated? Suppose that a firm serves locations 1 and 2 with leadtime demands—$\mathbf{LTD}_1$ and $\mathbf{LTD}_2$—that are statistically identically distributed but correlated. We represent the correlation of demand between the two locations with a correlation coefficient u. The mean of the total leadtime demand is thus

$$LTD + LTD = 2\,LTD$$

Its variance is

$$\sigma^2_{LTD} + \sigma^2_{LTD} + 2u\sigma^2_{LTD} = 2(1 + u)\sigma^2_{LTD}$$

Therefore, the total safety inventory in the centralized system is

$$I^c_{safety} = z \times \sqrt{2(1 + u)}\sigma_{LTD} \qquad \textbf{(Equation 7.10)}$$

The total safety inventory required in the decentralized system is

$$I^d_{safety} = 2 \times z \times \sigma_{LTD}$$

Therefore, the safety inventory in the two-location decentralized system is larger than in the centralized system by a factor of

$$\sqrt{\frac{2}{(1 + u)}}$$

When the demands of the two locations are independent, the correlation coefficient $u = 0$ and the safety inventory in the decentralized system is larger by a factor of $\sqrt{2}$, as

discussed earlier. The advantage of a centralized system increases as the demands on the two locations become negatively correlated. The advantage of a centralized system, however, diminishes as the demands in the two locations become positively correlated. In fact, if demand is perfectly positively correlated (i.e., $u = 1$), centralization offers no benefits in the reduction of safety inventory. The benefits of economies of scale discussed in Chapter 6, however, remain.

Disadvantages of Centralization If centralization of stocks reduces inventory, why doesn't every firm practice it? In addition to inventory costs, at least two other factors need to be considered in making the decision. First is the *response time* to the customer, which measures the time taken to satisfy the customer demand. Second is the *shipment cost* of sending the goods to the customer from the warehouse. In the previous analysis, we assumed that both centralized and decentralized operations have identical response times and identical shipment costs. Hence, inventory costs remained the only determining factor. In practice, a centralized location will typically be farther away from some customer locations than are some decentralized locations; centralization may entail longer response times when units must be shipped to more distant customers. It may also be more expensive to transport products to customers located at vastly different distances from the central location. If, in addition, response time and/or shipping costs increase such that overall demand decreases, then serving every customer from a single facility may not be optimal. In such situations, decentralized locations may improve response times and service levels. With decentralized locations, proximity to customers offers the opportunity to better understand their needs and develop closer relationships.

In addition, companies need to be aware of the cultural, linguistic, and regulatory barriers in some parts of the world (e.g., Europe and Asia) that may inhibit them in centralizing their operations. In fact, GE Lighting's decision, as illustrated at the beginning of this chapter, demonstrates that companies indeed consider these factors important. The trade-offs involved in centralization versus decentralization led GE Lighting to build a hybrid network where some parts of Europe were served by a single distribution facility, whereas others were served by local warehouses (Harps, 2000).

7.5.2 Principle of Aggregation and Pooling Inventory

It is important to stress that the inventory benefits outlined in the previous subsection result from the statistical principle called the **principle of aggregation**, which states that *the standard deviation of the sum of random variables is less than the sum of the individual standard deviations.* In all the examples discussed in Section 7.5.1, total inventory is physically located at a central location to enable the seller to aggregate demand across various regions. Physical consolidation, however, is not essential. As long as *available inventory is shared among various sources of demand—a practice known as* **pooling inventory**. Whenever this practice is taken into account for inventory placement, we achieve the benefits of aggregation. As the following examples indicate, the concept of pooling inventory can be applied in various ways other than physical centralization.

Virtual Centralization Consider a distribution system with warehouses in two locations, A and B. Each location carries some safety stock to provide a given level of service. Suppose now that at a given time, demand for a product in location A exceeds the available local stock. The product, however, is available at location B. Customer demand at location A can then be satisfied with stock at location B. Likewise, if at any time location B is out of stock on an item that is available at location A, the product can

be shipped to location B to satisfy customer demand. To accommodate this option, however, a system must satisfy two criteria:

1. Information about product demand and availability must be available at both locations.
2. Shipping the product from one location, B, to a customer at another location, A, must be fast and cost effective.

If these two requirements are met and correlation of demand is less than one, pooling is effective—inventory at any location can be shared by demands at all other locations. Because pooling is achieved by keeping the inventories at decentralized locations instead of physically consolidating them at one location, we call it virtual centralization. Formally, **virtual centralization** is *a system in which inventory pooling in a network of locations is facilitated using information regarding availability of goods and subsequent transshipment of goods between locations to satisfy demand.*

Machine-tool builder Okuma America Corporation, a subsidiary of Japan's Okuma Corporation, is an example of a company that is moving its distribution network toward virtual centralization. Each of its 46 distributors in North and South America has access to Okumalink, a shared information technology system that provides information about the location and availability of machine tools stored in Okuma warehouses in Charlotte, North Carolina, and in Japan. Okumalink is currently being upgraded to allow channel members to connect with one other directly, thereby facilitating intra channel exchanges of products and parts (Narus & Anderson, 1996). Similarly, in the event of a stockout, orders placed to a distribution center in W.W. Grainger, a large industrial distributor, may be fulfilled from another distribution center in the network that has available stock.

Specialization A firm may have several warehouses, each of which stocks several products. Safety inventory for each product, however, may be allocated to a particular warehouse that specializes in that product. Even though there are several warehouses, there is for each product only one specialized warehouse that carries the entire safety inventory. Each warehouse effectively pools with all the others the inventory for the product in which it specializes.

This system is particularly useful when the local demand that each warehouse serves is more or less unique to the product. For example, suppose there are two warehouses, one each at locations A and B. Suppose inventory consists of two products, P1 and P2. In addition, suppose a large fraction of demand at location A is for product P1 and a large fraction of that at location B is for product P2. Then location A's warehouse may be specialized to carry all the safety stock for product P1, and location B may be specialized for product P2. If location B (or A) requires any units of P1 (or P2), it could be shipped from location A (or B). Under this arrangement, safety inventory for each product is reduced because each inventory is now centralized at one location. Furthermore, because centralization is based on the local demand patterns, response times and shipping costs are also less than they would be if all products were physically centralized at one warehouse.

Component Commonality Our examples thus far have focused on pooling efficiency by means of aggregating demand across multiple geographic locations. The concept of pooling can also be exploited when aggregating demand across various products. Consider a computer manufacturer (such as Dell, Hewlett Packard, or Apple) that typically offers a wide range of models. Although models vary considerably, a few common components are used across product lines, such as similar central processing units or DVD or CD-RW drives.

To offer variety, firms have a few options. They can, for instance, *produce in anticipation of product demand*—a **make-to-stock** strategy. To provide a desired service level,

the firm would then need sufficient safety inventory of each final product. Conversely, the firm may decide to *produce in response to customers orders*—a **make-to-order** strategy. Under this strategy, a firm keeps its entire inventory in components and builds the final product as and when customers place orders. To determine the safety inventory of those components common to various product lines, the firm aggregates demand for the products that share specific components. Component commonality thus allows the firm to reduce inventory investment while maintaining the same level of service and offering product variety.

Risk pooling of common-component demand across various products is akin to the practice of physical centralization that we described earlier. Safety inventory of common components will be much lower than the safety inventory of unique components stored separately for different finished products. In addition, in a make-to-order situation, holding costs will be less because inventory of components has accumulated no added value. There is, however, at least one key drawback. In a make-to-order situation, the customer must wait for the firm to produce the product, whereas the make-to-stock product is available for immediate consumption. Therefore, if flow times in production can be shortened until they are shorter in duration than the wait that the consumer is willing to endure, then a make-to-order strategy has significant benefits.

Product Substitution Often, one product can substitute to fill excess demand for another. The ability to provide substitute products improves the effective level of service by pooling safety inventory across multiple products. Substitution, therefore, reduces the level of safety stock needed for a given level of customer service. To exploit this, however, a firm needs to gather information on substitution patterns. Retailers often place substitute products next to each other on the retail shelf. By learning the substitution behavior of consumers within a product category, a retailer can better optimize inventory level on the shelf.

7.6 SHORTENING THE FORECAST HORIZON THROUGH POSTPONEMENT

As noted in Section 7.1, forecasts further into the future tend to be less accurate than those of more imminent events. Quite simply, as time passes, we get better information and so can make better predictions. Because shorter-range forecasts are more accurate, inventory-planning decisions will be more effective if supply is postponed closer to the point of actual demand.

Postponement (or Delayed Differentiation) Consider a garment manufacturer who makes blue, green, and red T-shirts. The firm is considering two alternate manufacturing processes:

1. Process A calls first for coloring the fabric, which takes one week, and then assembling the T-shirt, which takes another week.
2. Process B calls first for assembling T-shirts from white fabric, which also takes one week, and then coloring the assembled shirts—a process that, as in process A, takes one week.

Both processes, therefore, take a total of two weeks. Does one have any advantage over the other? With process A, the manufacturer must forecast demand for T-shirts in every color that will sell in two weeks. Although total flow time per T-shirt is the same under both processes A and B, by reversing the assembly and dyeing stages, process B has essentially postponed the color differentiation until one week closer to the time of sale. *The practice of reorganizing a process in order to delay the differentiation of a generic product to specific end-products closer to the time of sale* is called **postponement** or

delayed differentiation. Because it is easier to more accurately forecast demand for different colored T-shirts for next week than demand for the week after next, process B will entail less safety inventory of colored T-shirts than process A.

Process B also has another advantage. In deciding the number of white T-shirts to assemble in the first phase, the manufacturer can make an aggregate forecast across all colors (as discussed, aggregation reduces variability). Process B, then, boasts reduced variability for two reasons:

1. It aggregates demands by color in the first (assembly) phase.
2. It requires shorter-range forecasts of individual T-shirts needed by color in the second (dyeing) phase.

Both result in less demand variability and hence require less total safety inventory.

Clothing maker Benetton (Signorelli & Heskett, 1989) was an early pioneer in postponement strategies. Another company that has found this particular process innovation beneficial is Hewlett Packard (HP), which builds Deskjet printers for worldwide sales (Kopczak & Lee, 2001). For example, one major difference between printers destined for North America and those bound for Europe is their power-supply rating. Initially, the HP Deskjet printer was designed to include a specific power supply (110 or 220 volts) in the assembly process—the plant would make printers specific to each geographical location. But in rethinking its distribution system, HP redesigned the printer so that the power-supply module could be installed at the very end of the production process (in fact, it is installed by the distributor). Thus, the plant was producing a generic printer and postponing differentiation until the distribution stage. The more recent HP Deskjet printers carry the postponement concept even further. These printers can be used as either color or black-and-white printers simply by inserting the appropriate cartridge. Because this customization process is actually performed by the consumer, HP has no need to forecast separate demand for color and black-and-white printers. A similar postponement strategy was adopted by Dade Behring (DB), an industry leader in clinical diagnostic equipment and reagents. DB manufactures high-end diagnostic instruments ranging in prices from $20,000-$200,000. It adopted a two-pronged postponement strategy. The first postponement point involved product redesign to adopt a universal power supply instead of dedicated 110V or 220V power supply. In addition, to adapt to the European In Vitro Diagnostics Directive (IVDD) that called for medical and diagnostic equipment to come packaged with local language manuals and labeling, DB used its distribution center to package language specific manuals and also developed flexible language capability within the equipment operating software (Rietze, 2006).

7.7 PERIODIC REVIEW POLICY

The discussion thus far has focused on determination of safety stock policies when the firm has a capability to monitor inventory continuously, so as to trigger a reorder as soon as the inventory position reaches a predetermined point. As mentioned in Chapter 6, Section 6.7, firms may for several reasons choose to operate their inventory system (review and reorder) in a periodic fashion. There we showed that firms following a **periodic review policy** will periodically reorder to raise the inventory position to a fixed target called the **order upto level** (*OUL*). We determined that in the absence of demand uncertainty the order upto level for a firm is given by $OUL = R \times (T_r + L)$, where R is the flow rate, T_r is the **review period** and L is the replenishment lead time. When faced with demand uncertainty, how should this order upto level be adjusted to achieve a target service level? Because under a periodic review system, the target inventory level needs to account for demand dynamics during the length of the review period as well as the replenishment lead time, using arguments analogous to the continuous

review system articulated in section 7.2.2, it should be evident that we need to carry some safety stock to buffer against uncertainty in demand over the review period *and* the replenishment lead time so as to achieve a cycle service level of greater than 50%. Notice that for the periodic review case, the time interval of interest to measure the cycle service level (see Section 7.2.1) is review period plus replenishment lead time.

To estimate the safety inventory, we need to first estimate the standard deviation of demand during review period and replenishment lead time. In section 7.4.1, we illustrated how to estimate the standard deviation of leadtime demand. We now have to estimate standard deviation of demand over $(T_r + L)$ periods. Using similar arguments, we can show that standard deviation of demand during a review period and lead time, denoted by σ_{RLTD} will be

$$\sigma_{RLTD} = \sqrt{T_r + L} \times \sigma_R \qquad \textbf{(Equation 7.11)}$$

Safety inventory to achieve a given service level SL can then be articulated as a multiple z of the standard deviation during review period and leadtime demand. That is,

$$I_{safety} = z \times \sigma_{RLTD} \qquad \textbf{(Equation 7.12)}$$

Finally, the order upto level is given by

$$\begin{aligned} OUL &= \text{Average demand during } (T_r + L) + \text{Safety Inventory} \\ &= R \times (T_r + L) + I_{safety} \end{aligned} \qquad \textbf{(Equation 7.13)}$$

Example 7.11 below illustrates the calculation.

EXAMPLE 7.11

Consider the GE Lighting warehouse near Paris. The throughput rate of lamps is, say, 2,000 units per day. Suppose the standard deviation of daily demand is 1581 units and the warehouse manager follows a periodic review policy with a review period of 14 days. Whenever the manager places an order, the replenishment is received in 10 days. What should be the order upto level to achieve a service level of 95%?

The average demand during review period of $T_r = 14$ days and a replenishment lead time of $L = 10$ days is

$$R \times (T_r + L) = 2000 \times (14 + 10) = 48{,}000 \text{ units}$$

The standard deviation of demand for the same duration is

$$\sigma_{RLTD} = \sqrt{T_r + L} \times \sigma_R = \sqrt{14 + 10} \times 1581 = 7{,}745$$

The z-value to achieve a service level of 95% is given by:

$$z = \text{NORMSINV } (0.95) = 1.65$$

and safety inventory

$$I_{safety} = z \times \sigma_{RLTD} = 1.65 \times 7{,}745 = 12{,}779$$

We then estimate the order up to level as

$$OUL = R \times (T_r + L) + I_{safety} = 48{,}000 + 12{,}779 = 60{,}779$$

Thus the manager will set a target inventory level of 60,779 units and every 14 days place a reorder to bring the inventory position to this level. While the exact order quantity will fluctuate between order cycles, the average order size will be the average demand over 14 days and equal to 28,000 units. Therefore, the cycle inventory, $I_{cycle} = 28{,}000/2 = 14{,}000$. A sample inventory profile is illustrated in Figure 7.5.

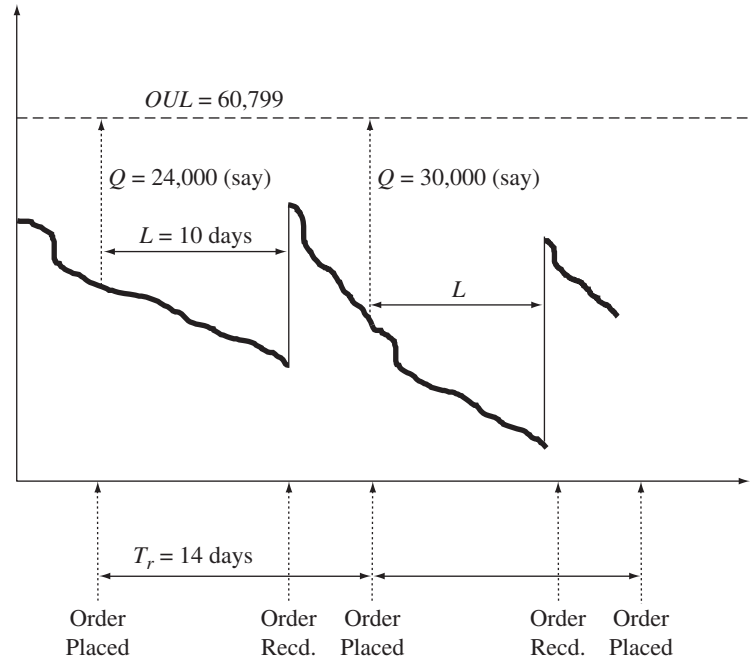


FIGURE 7.5 Inventory Profile for a Periodic Review Policy

It is useful to compare the safety inventory between periodic review and continuous review system. Recall that in a continuous review system, safety inventory needs to buffer against demand uncertainty only during lead time and will be equal to

$$z \times \sigma_{LTD} = 1.65 \times \sqrt{10} \times 1{,}581 = 8{,}249$$

which is significantly less than the 12,779 units of safety inventory necessary to achieve the same service level in a periodic review system. Therefore, we conclude that the inability to monitor inventory in real time and take action accordingly results in additional inventory costs. Investment in an information technology system and efforts to reduce other costs related to periodic ordering will allow a process manager to reduce the review period resulting in reduced cycle and safety inventory.

7.8 LEVERS FOR REDUCING SAFETY INVENTORY

In this chapter, we first recognized the role of uncertainty and variability in process inflows and outflows and introduced the notion of *safety inventory* as a buffer against uncertainty in supply and/or demand.

We can identify the following levers for reducing flow variability and the required safety inventory (and thus flow time):

1. Reduce demand variability through improved forecasting
2. Reduce replenishment lead time
3. Reduce review period length
4. Reduce variability in replenishment lead time
5. Pool safety inventory for multiple locations or products, whether through physical or virtual centralization or specialization or some combination thereof
6. Exploit product substitution
7. Use common components
8. Postpone product-differentiation processing until closer to the point of actual demand

Summary

Firms carry *safety inventory* of inputs and outputs as protection against possible stockouts resulting from unexpected supply shortages and demand surges. The goal is to ensure that flow units are available to meet the company's production needs and customer requirements despite supply and demand uncertainty. The probability that flow units will be available to satisfy customer requirements is called service level, which measures the degree of stockout protection provided by a given amount of safety inventory. The higher the level of safety inventory, the higher the level of service provided. The optimal service level balances the net marginal benefit of each additional unit with its net marginal cost as given by the newsvendor formula.

Both the service level provided and the safety inventory required depend on *variability* in flow rates—reducing variability increases the service level that is provided by a given amount of safety inventory and decreases the amount of safety inventory that is necessary to provide a given level of service. Variability of flow rates in turn depend on the forecast errors, length of the review period and the mean and variability of replenishment lead times. Therefore, better forecasting, shorter review periods, and fast and reliable suppliers are key to reducing investment in safety inventory. Pooling inventories to satisfy aggregate demand across multiple regions can also effectively decrease variability of flow rates and hence safety inventory. The square root law suggests that when demands are independent across regions, pooling reduces safety inventory by a factor of the square root of the number of locations aggregated. The aggregation principle can be operationalized in several other ways including virtual centralization, specialization, and component commonality. Finally, since forecast errors decrease closer to the point of sale, strategies to postpone critical decisions on differentiation will allow a firm to reduce safety inventory without sacrificing service level.

Key Equations and Symbols

(Equation 7.1) $ROP = LTD + I_{safety}$

(Equation 7.2) $I = I_{cycle} + I_{safety} = Q/2 + I_{safety}$

(Equation 7.3) $SL = \text{Prob}(\mathbf{LTD} \leq ROP)$
(continuous review system)

(Equation 7.4) $I_{safety} = z \times \sigma_{LTD}$
(*continuous review system*)

(Equation 7.5) *Newsvendor formula*:

$$SL^* = \text{Prob}(\mathbf{R} \leq Q^*) = \frac{MB}{MB + MC}$$

(Equation 7.6) $LTD = L \times R$

(Equation 7.7) $\sigma_{LTD} = \sqrt{L} \times \sigma_R$ (*fixed leadtime case*)

(Equation 7.8) $\sigma_{LTD} = \sqrt{L\sigma_R^2 + R^2\sigma_L^2}$
(*variable leadtime case*)

(Equation 7.9) $I^c_{safety} = z \times \sqrt{N}\sigma_{LTD}$
(*continuous review, N locations , no correlation*)

(Equation 7.10) $I^c_{safety} = z \times \sqrt{2(1+\rho)}\sigma_{LTD}$
(*continuous review, two locations with correlation*)

(Equation 7.11) $\sigma_{RLTD} = \sqrt{T_r + L} \times \sigma_R$
(*fixed review period and leadtime case*)

(Equation 7.12) $I_{safety} = z \times \sigma_{RLTD}$ (*periodic review system*)

(Equation 7.13) $OUL = R \times (T_r + L) + I_{safety}$
(*periodic review system*; I_{safety} *from Equation 7.12*)

where
ROP = Reorder point
LTD = Average leadtime demand
I_{safety} = Safety inventory
Q = Order size
I = Average inventory
SL = Service level
z = Service level factor
σ_{LTD} = Standard deviation of leadtime demand
SL^* = Optimal service level
MB = Net marginal benefit from each additional unit
MC = Net marginal cost of each additional unit
R = Average demand rate
σ_R = Standard deviation of period demand
L = Average replenishment lead time
σ_L = Standard deviation of replenishment lead time
I^c_{safety} = Safety inventory upon centralization
u = Demand correlation
T_r = Review period
OUL = Order upto Level

Key Terms

- Aggregation
- Backlogged
- Causal models
- Continuous review, reorder point policy
- Cycle service level
- Delayed differentiation
- Fill rate
- Forecasting
- Leadtime demand
- Make-to-order
- Make-to-stock
- Marginal analysis
- Net marginal benefit
- Net marginal cost
- Newsvendor problem
- Order upto level
- Periodic Review Policy
- Physical centralization
- Pooling inventory
- Postponement
- Principle of aggregation
- Review period
- Safety inventory
- Safety stock
- Square root law
- Time-series analyses
- Virtual centralization

Discussion Questions

7.1 What is the role of safety inventory?

7.2 Discuss the pros and cons of different ways to measure service level.

7.3 How does service level impact the level of safety inventory?

7.4 Consider two products with the same margins but with different salvage values. Which product should have a higher service level, and why?

7.5 If the quality of goods provided by suppliers is identical, purchasing goods based on lowest price is the best strategy. Discuss.

7.6 It takes the same amount of inventory to operate a single warehouse system as a four warehouse distribution network. True or false? Explain.

7.7 Going online allows a firm to supply online orders from a centralized location rather than using many retail outlets because customers are willing to wait a little for the online order to be delivered. Do you think that the inventory benefits of this centralization will be higher for staple grocery products like cereal and pasta or for products like music CDs and DVDs? Explain.

7.8 In the early days of paint manufacturing, manufacturers of paint used to produce paint of appropriate colors and sizes to be sold in retail stores. Today, consumers go to retail stores and select the color they wish, and the retailer mixes the pigment into a base paint to make the chosen color. Discuss what impact, if any, this strategy has on safety inventories.

7.9 Discuss how the inability to monitor and reorder inventory on a real-time basis impacts safety inventory.

Exercises

7.1 MassPC Inc. produces a 4-week supply of its PC Pal model when stock on hand drops to 500 units. It takes 1 week to produce a batch. Factory orders average 400 units per week, and standard deviation of forecast errors is estimated at 125 units.

a. What level of customer service is MassPC providing to its distributors in terms of stock availability?

b. MassPC wants to improve customer service to 80%, 90%, 95%, and 99%. How will such improvements affect the company's reorder policy and its annual costs?

*__7.2__ Weekly demand for DVD-Rs at a retailer is normally distributed with a mean of 1,000 boxes and a standard deviation of 150. Currently, the store places paper orders faxed to the supplier. Assume 50 working weeks in a year and the following data:

- leadtime for replenishment of an order is 4 weeks.
- Fixed cost (ordering and transportation) per order is $100.
- Each box of DVD-Rs costs $9.99.
- Annual holding cost is 25% of average inventory value.
- The retailer currently orders 20,000 DVD-Rs when stock on hand reaches 4,200.

a. Currently, how long, on average, does a box of DVD-Rs spend in the store? What is the annual ordering and holding cost under such a policy?

b. Assuming that the retailer wants the probability of stocking out in a cycle to be no more than 5%, recommend an optimal inventory policy (a policy regarding order quantity and safety stock). Under your recommended policy, how long, on average, would a box of DVD-Rs spend in the store?

c. Claiming that it will lower lead time to 1 week, the supplier is trying to persuade the retailer to adopt an electronic procurement system. In terms of costs and flow times, what benefits can the retailer expect to realize by adopting the electronic procurement system?

7.3 The Home and Garden (HG) chain of superstores imports decorative planters from Italy. Weekly demand for planters averages 1,500 with a standard deviation of 800. Each planter costs $10. HG incurs a holding cost of 25% per year to carry inventory. HG has an opportunity to set up a superstore in the Phoenix region. Each order shipped from Italy incurs a fixed transportation and delivery cost of $10,000. Consider 52 weeks in the year.

a. Determine the optimal order quantity of planters for HG.

b. If the delivery lead time from Italy is 4 weeks and HG wants to provide its customers a cycle service level of 90%, how much safety stock should it carry?

c. Fastship is a new shipping company that promises to reduce the delivery lead time for planters from 4 weeks to 1 week using a faster ship and expedited customs clearance. Using Fastship will add $0.2 to the cost of each planter. Should HG go with Fastship? Why or why not? Quantify the impact of the change.

7.4 Johnson Electronics sells electrical and electronic components through catalogs. Catalogs are printed once every two years. Each printing run incurs a fixed cost of $25,000, with a variable production cost of $5 per catalog. Annual demand for catalogs is estimated to be normally distributed with a mean of 16,000 and standard deviation of 4,000. Data indicates that, on average, each customer ordering a catalog generates a profit of $35 from sales. Assuming that Johnson wants only one printing run in each two-year cycle, how many catalogs should be printed in each run?

***7.5** As owner of Catch-of-the-Day Fish Shop, you can purchase fresh fish at $18 per crate each morning from the Walton Fish Market. During the day, you sell crates of fish to local restaurants for $120 each. Coupled with the perishable nature of your product, your integrity as a quality supplier requires you to dispose of each unsold crate at the end of the day. Your cost of disposal is $2 per crate. You have a problem, however, because you do not know how many crates your customers will order each day. To address this problem, you have collected the several days' worth of demand data shown in Table 7.5. You now want to determine the optimal number of crates you should purchase each morning.

7.6 The residents of Bucktown, Illinois, place their trash at the curb each Wednesday morning to be picked up by municipal crews. Experience shows that the total amount of trash put out has a normal distribution with a mean of 35 tons and a standard deviation of 9 tons. Crews of full-time city employees assigned to trash collection collect trash. Each crew can collect 5 tons of trash per working day. The city has plenty of trucks of the kind used for trash collection. The marginal cost of operating one trash collection crew for one working day, including both personnel-related costs and truck-related costs, is reckoned at $625. Whatever trash remains at the end of the work day *must* be collected that evening by an outside contractor who charges $650 per ton.

How many crews should the city assign to trash collection? For simplicity, treat the number of crews as a continuous variable.

7.7 Northwest Airlines runs daily flights from Detroit to Amsterdam. They face a fixed cost of $70,000 for each flight independent of the actual number of passengers on the plane. There are 310 seats available on a plane. One-way tickets generate revenues of $600 apiece when used but are fully refundable if not used. On a typical weekday, the airline estimates that the number of no-shows will range between 0 and 20; all intermediate values are equally likely.

By law, an airline is allowed to overbook flights, but must give compensation of $250 to all ticketed passengers not allowed to board. In addition, it must provide those passengers with alternative transportation on another carrier (the cost of providing the alternative transportation just wipes out the $600 revenue). How many tickets should Northwest book on its flight from Detroit to Amsterdam?

***7.8** A mail-order firm has four regional warehouses. Demand at each warehouse is normally distributed with a mean of 10,000 per week and a standard deviation of 2,000. Annual holding cost is 25%, and each unit of product costs the company $10. Each order incurs an ordering cost of $1,000 (primarily from fixed transportation costs), and lead time is 1 week. The company wants the probability of stocking out in a flow to be no more than 5%. Assume 50 working weeks in a year.

a. Assuming that each warehouse operates independently, what should be the ordering policy at each warehouse? How much safety stock does each warehouse hold? How much average inventory is held (at all four warehouses combined) and at what annual cost? On average, how long does a unit of product spend in the warehouse before being sold?

Table 7.5 Demand at Catch-of-the-Day Fish Shop

Demand	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frequency	0	0	0	1	3	2	5	1	6	7	6	8	5	4	1	3

b. Assume that the firm has centralized all inventories in a single warehouse and that the probability of stocking out in a cycle can still be no more than 5%. Ideally, how much average inventory can the company now expect to hold and at what cost? In this case, how long will a unit spend in the warehouse before being sold?

7.9 Hi-Tek is a retailer of computer equipment in the greater Chicago region with four retail outlets. Currently, each outlet manages its ordering independently. Demand at each retail outlet averages 4,000 units per week. Each unit costs $200, and Hi-Tek has an annual holding cost of 20%. The fixed cost of each order (administrative + transportation) is $900. Assume 50 weeks in a year.

a. Given that each outlet orders independently and gets its own delivery, determine the optimal order size at each outlet.

b. On average, how long (in weeks) does each unit spend in the Hi-Tek system before being sold?

c. Hi-Tek is thinking of centralizing purchasing (for all four outlets). In this setting, Hi-Tek will place a single order (for all outlets) with the supplier. The supplier will deliver the order on a common truck to a transit point. Since individual requirements are identical across outlets, the total order is split equally and shipped to the retailers from this transit point. This entire operation will increase the fixed cost of placing an order to $1,800. If Hi-Tek manages ordering optimally, determine the average inventory across all four outlets in the new Hi-Tek system.

*__7.10__ Daily consumption of fasteners at Boeing commercial airplane manufacturing facility is normally distributed with a mean of 1,000 and a standard deviation of 150. Each fastener costs $2. Boeing reviews its inventory every 2 weeks and places an order to bring the inventory position of fasteners to a target level. leadtime for replenishment of an order is 1 week. Annual holding cost is 25% of unit cost.

a. Assuming that Boeing wants to keep a 98% in-stock probability, determine the target order upto level and the resulting cycle and safety stock. What is the total inventory holding cost of following this policy?

b. There is a proposal to institute process improvements such that inventory could be reviewed and orders placed every week. Determine the total savings in inventory of this process change.

7.11 Reconsider Exercise 7.8. Now suppose that the mail-order firm follows a periodic review policy with a review period of 2 weeks. Recall that the firm has four regional warehouses with demand at each warehouse that is normally distributed with a mean of 10,000 per week and a standard deviation of 2,000. Further, annual holding cost is 25%, and each unit of product costs the company $10. replenishment lead time is 1 week. The company wants a service level of 95%. Assume 50 working weeks in a year.

a. Assuming that each warehouse operates independently, what should be the ordering policy at each warehouse? How much safety stock does each warehouse hold? How much average inventory is held (at all four warehouses combined) and at what annual cost? On average, how long does a unit of product spend in the warehouse before being sold?

b. Assume that the firm has centralized all inventories in a single warehouse and that the target service level is still 95%. Ideally, how much average inventory can the company now expect to hold and at what cost? In this case, how long will a unit spend in the warehouse before being sold?

Selected Bibliography

Jacobs, F. R., and R. Chase. *Operations and Supply Chain Management*. 13th ed. New York: McGraw-Hill/Irwin, 2010.

Chopra, S., and P. Miendl. *Supply Chain Management: Strategy Planning and Operations*. 4th ed. Upper Saddle River, N.J.: Prentice Hall, 2009.

Gruen, T. W., D. S. Corsten, and S. Bharadwaj. *Retail Out-of-Stocks: A Worldwide Examination of Extent, Causes, and Consumer Responses*. Washington, D.C.: Grocery Manufacturers of America, Industry Affairs Group, 2002.

Harps, L. H. "Pan-European Distribution." *Logistics Management* 39, (February 2000): 2.

Kopczak, L., and H. L. Lee. *Hewlett Packard: Deskjet Printer Supply Chain*. Graduate School of Business, Stanford University, Case Study #GS-3A. 2001

Nahmias, S. *Production and Operations Analysis*. 6th ed. Homewood, Ill.: McGraw-Hill/Irwin, 2008.

Narus, J., and J. S. Anderson. "Rethinking Distribution: Adaptive Channels." *Harvard Business Review* (July-August 1996): 112–120.

Rietze, S.. Case Studies of Postponement in Supply Chain. M.S. Thesis, MIT, Cambridge, MA, 2006.

Ross, S. *Introduction to Probability Models*. New York: Academic Press, 1972.

Signorelli, S., and J. L. Heskett.. *Benetton (A)*. 9-685-014. Cambridge, Mass.: Harvard Business School Publishing, 1989. 1–20.

APPENDIX

Calculating Service Level for a Given Safety Inventory

The service level for a given ROP is given by

$$SL = \text{Prob}(\mathbf{LTD} \leq ROP)$$

To calculate *SL*, recall first that if **LTD** is normally distributed with mean *LTD* and standard deviation σ_{LTD}, then

$$\mathbf{Z} = (\mathbf{LTD} - LTD)/\sigma_{LTD}$$

is also normally distributed with mean 0 and standard deviation 1 and is known as the standard normal random variable.

Furthermore, a given level of safety inventory, I_{safety}, can be measured as a multiple, z, of the standard deviation σ_{LTD} of **LTD**. Thus, we can say the following:

$$I_{safety} = z \times \sigma_{LTD}$$

Using the fact that $ROP = I_{safety} + LTD$, we write

$$z = \frac{ROP - LTD}{\sigma_{LTD}}$$

Therefore, we can say the following:

$$\begin{aligned} SL &= \text{Prob}(\mathbf{LTD} \leq ROP) \\ &= \text{Prob}\left(\frac{\mathbf{LTD} - LTD}{\sigma_{LTD}} \leq \frac{ROP - LTD}{\sigma_{LTD}}\right) \\ &= \text{Prob}(\mathbf{Z} \leq z) \end{aligned}$$

Thus we show the relationship between service level and the z-value; specifically, service level is the area under the standard normal density curve below the z-value, where the z-value depends on the safety inventory.

Chapter

8

Managing Flow Variability: Safety Capacity

INTRODUCTION

L. L. Bean is a mail-order retailer of outdoor gear and casual apparel, renowned for high quality customer service. In addition to online and brick and mortar stores, they are a well-known mail-order catalog house served by call centers throughout Maine. Customers call a toll-free number to place their orders with customer service representatives (CSRs) who provide them product information, take down their orders, and set up shipping and billing schedules. Each caller gets connected to a CSR, or is put on hold to wait for one, or gets a busy signal. Customers who get served by CSRs generate profit for the retailer and exit the process. Customers who experience long waits on hold may decide to hang up, while those who get busy signals do not even get into the system. In either case, they may call back later, or they may contact and buy from a competitor, costing L. L. Bean profit from lost sales. Figure 8.1 shows the call center process that customers go through.

Lately, L. L. Bean's customers have been complaining about frequently receiving busy signals and spending excessive times on hold at one of their call centers in Bangor. Since the call center is the retailer's main source of revenue, the management is concerned about its performance and the effect on the retailer's bottom line. They have asked their performance

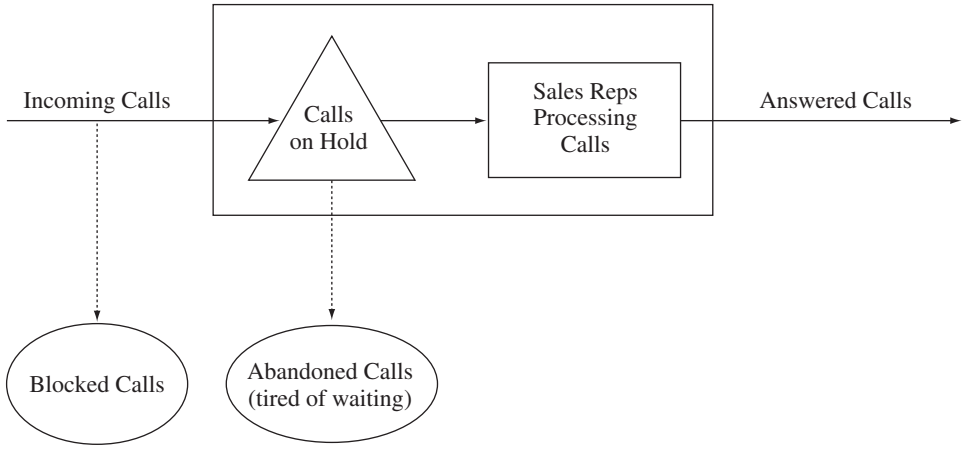


FIGURE 8.1 The Service Call Center at L. L. Bean

improvement team (PIT) to identify the causes of customer delays and defections, and make necessary changes in the call center design and operation to improve its performance, which would lead to greater customer satisfaction and higher sales.

In this chapter, we study how to model and analyze such order fulfillment processes in order to gain managerial insights into their performance measurement and improvement. As we will see, the fundamental problem in managing these operations involves matching their processing capacity with variable demand.

In Chapter 7, we studied the role of inventory in satisfying variable demand to ensure product availability. We saw that carrying sufficient safety inventory permits meeting unanticipated demand to provide a desired level of customer service. Optimal amount of safety inventory balances the cost of carrying it against the benefit of improved product availability. Throughout our analysis, the implicit assumption was that the product being supplied can be produced and inventoried in advance of the actual demand. We were thus dealing with the problem of matching supply with demand in make-to-stock operations.

However, many businesses involve make-to-order operations such as job shops and service facilities (like the L. L. Bean's call center) wherein customer orders cannot be processed and inventoried ahead of time. In fact, order processing may not even begin until after the order is received. Thus, for example, we cannot inventory already completed telephone calls in a call center, or already performed surgeries in a hospital, or already served meals in a restaurant! Without the benefit of inventory of finished orders, the process manager must keep sufficient capacity to process orders as they come in. These orders now become inflow units that may have to wait before being processed by the available resources. The queue of waiting orders then becomes the inventory of flow units, and the resulting delay increases the order flow time. The cost of holding this inventory of unprocessed orders arises from the customer dissatisfaction with delays in getting their orders filled. The process manager must balance these costs of orders waiting against the cost of process improvement required to reduce it. In this chapter, we will identify the causes of these queues and delays, and suggest appropriate managerial actions to reduce them.

To be sure, in practice, most business processes involve a combination of make-to-stock and make-to-order operations. For example, a computer manufacturer, such as Dell, may produce and stock partially assembled computers ahead of the actual demand, but finish their final assembly according to customer specifications only after receiving a specific order. Similarly, a fast-food restaurant, such as McDonald's, may prepare partially

assembled sandwiches before the rush hour, and complete the orders to individual requirements only after customers come in. In such hybrid operations, the process manager can manipulate *both* inventory (of partially finished products) and capacity (to complete the processing) in an attempt to match supply with demand. However, to understand these two levers more clearly, we have chosen to isolate them by focusing first on inventory alone in purely make-to-stock operations in the preceding chapter, and on capacity alone in purely make-to-order operations in this chapter. Moreover, for concreteness, we will address service operations, although the concepts, methods, and conclusions are equally applicable to manufacturing in job shops and other make-to-order operations.

In Section 8.1, we describe a typical service process, outline key measures of its performance in terms of queues and delays, and identify insufficient capacity as the obvious reason for unsatisfactory performance. In Section 8.2, we study how variability in customer arrival and **processing time** is also responsible for these queues and delays, even with sufficient processing capacity. In Section 8.3, we quantify capacity utilization and process variability as the two main drivers of process performance. In Section 8.4, we study the economics of capacity investment decisions to reduce delays. In Section 8.5, we consider the effect of limited buffer capacity in blocking of arrivals and their abandonment after long delays. In Section 8.6, we consider the worst-case—rather than the average— performance of a service process and the turn-around time that we can promise with a certain level of confidence. In Section 8.7, we consider the effect of pooling capacity across homogeneous arrivals, segregating and prioritizing heterogeneous arrivals, and the role of resource flexibility and specialization in improving the overall performance. In Section 8.8, we summarize practical managerial actions to improve the process performance by increasing and pooling capacity, reducing variability in arrivals and processing, and synchronizing the available capacity with demand. Finally, in Section 8.9, we discuss how customer perceptions and expectations can be managed to mitigate the adverse effect of delays and congestion on customer satisfaction. We conclude the chapter by summarizing the key levers for improving process performance in terms of reduced queues and delays.

8.1 SERVICE PROCESS AND ITS PERFORMANCE

Recall that any business process is defined by its inputs, outputs, buffers, and resources. In the previous chapters, we emphasized flow rate, flow time, and inventory as the key operational metrics of process performance and identified various levers that drive these metrics. In the preceding chapter, we studied the role of inventories in matching supply with demand in make-to-stock operations. In this chapter, we consider this problem in make-to-order operations, such as job shops and service operations, where we cannot carry inventory of finished orders, and must carry sufficient capacity to process orders as they come in. These order fulfillment processes are characterized by (1) variability in order arrivals as well as in order processing, and (2) the use of safety capacity—rather than safety inventory—in dealing with this variability. Although our analysis is applicable to any make-to-order process, such as job shops, for concreteness, we will address service processes. Accordingly, we will specialize the general business process terminology to discuss service processes in more natural terms. However, for consistency of exposition, we will continue to use the same general notation as in the previous chapters.

8.1.1 Service Processes

In a service process, it is natural to refer to inflow of job orders as customer arrivals. Thus, customers at bank teller windows, drive-through restaurants, and supermarket checkouts, as well as passengers checking in at airline counters and patients taken to

hospital emergency rooms, are examples of customers arriving for service. Similarly, telephone calls ringing at call centers, information packets transmitted to data processing centers, planes landing at airports, and ships or trucks arriving at loading docks are also "customer" arrivals for processing. Finally, work orders submitted to job shops, projects undertaken by consulting firms, and cars brought to auto repair shops can all be viewed as "customers" that are processed by available resources.

A resource unit that processes a customer will be referred to as a server. Thus, airline agents, CSRs, airport runways, docking crews, doctors, and loan officers are "servers" that process passengers at an airline counter, telephone calls at a call center, airplanes landing or taking off, trucks at a loading dock, patients, and loan applications at a bank, respectively. If a server is not immediately available, the arriving customer must join and wait in a queue, which is simply the inventory of inflow units that accumulates in the input buffer due to insufficient processing capacity. (This is equivalent to customers having to wait for physical products because of insufficient inventory of finished goods in make-to-stock operations.) Thus, each stage of a passenger's travel from making reservations, checking in and obtaining a boarding pass, going through security, boarding plane, taking off, landing, reclaiming baggage, and renting a car involves waiting. Other examples where customers have to wait include banks, restaurants, supermarkets, hospitals, and amusement parks.

As we will see, customer waiting occurs due to insufficient processing capacity and variability in arrival and processing times. Variability in arrival times arises from the fact that the process manager has little control over the customer arrivals. Even with reservations, appointments, and careful scheduling, the actual times of customer arrivals may still display variability. Moreover, given the customized nature of make-to-order operations, their individual order processing times are also significantly more variable than in make-to-stock operations. For example, job shops typically handle a wide variety of orders, each with different processing requirements, so their processing times are also more variable than in flow shops which tend to focus on producing only a limited variety of products. Similarly, patients coming to a hospital emergency room have widely differing ailments, so the time required to treat each is very different as well. As we will see, this variability in customer arrival and processing times results in delays and queues of customers who have to wait for service. The process manager must balance the cost of reducing this variability and increasing processing capacity against the benefit of improved performance in terms of reduced delays and queues.

We will only consider a single-activity process, where each customer is processed by one server, and all tasks performed by that server are combined into a single activity. An example is the L. L. Bean call center, where each CSR handles all requirements of each caller. In a more complex processing network, multiple servers may process each customer, and processing may involve different activities performed in a sequence that is specified by a process flowchart. For example, processing an insurance claim involves several steps, such as policy verification, loss determination, and claim settlement, each involving different agents and information resources. Although in Chapters 4 and 5 we analyzed flow times and flow rates in such complex processing networks, we assumed away any variability in inflow and processing times. In this chapter, our focus will be on this variability and how to deal with it, so to keep the analysis manageable, we will suppress other process details and treat the entire process as a single activity performed by one resource unit. The resulting single-activity process model is easier to analyze, and still brings out most of the key managerial insights that we wish to gain.

A **single-phase service process** may *have multiple servers, each performing the same set of activities on one customer at a time*. With multiple servers, customer arrivals may form either a single queue that feeds into the pool of all servers or multiple queues, a separate one for each server. **Service order discipline** *is a specification of the sequence in*

which waiting customers are served. We will always assume that customers are served in the order of their arrival; that is, the service order discipline is first-come-first-served (FCFS). In a single-phase service process, there is no output buffer since customers exit the process as soon as they are served, as in the L. L. Bean call center where customers hang up as soon as they finish speaking with CSRs. Finally, to simplify the exposition, in this section we will assume that all customer arrivals can be accommodated in the input buffer and are eventually served; refinements in which some arrivals may be turned away due to limited waiting room capacity or some may leave the queue due to long waits will be discussed later in Section 8.7.

8.1.2 Service Process Attributes

In a typical single-phase service process, customers arrive, wait for service, and are processed by one of the available servers in the FCFS order. Customer flow through the service process is then characterized by its attributes that determine the demand for and supply of service.

- Arrival rate R_i is the average **inflow rate** of customer arrivals per unit time; it represents demand for service. The **interarrival time**, which is the average time between consecutive customer arrivals, is then $1/R_i$ time units.
- Service (or processing) time T_p is the average processing time required to serve a customer. Observe that for a single-phase service process, the processing time is the same as the activity time (Chapter 4) and the unit load (Chapter 5). The unit service rate (or **processing rate**) of a server is then $1/T_p$ customers per unit time; it is the processing capacity or *speed* of each server.
- All servers together form the server pool, and the number of servers in the pool will be denoted by c. Since each server processes one customer at a time, c is also the maximum number of customers that can be processed simultaneously; it represents the *scale* of the processing operation.
- **Service rate** (or process capacity) R_p is then the maximum rate at which customers can be processed by all of the servers in the server pool, so that

$$R_p = \frac{c}{T_p} \qquad \textbf{(Equation 8.1)}$$

which can be interpreted as *scale* × *speed.* Process capacity measures its total supply of service, in terms of the maximum number of customers that can be processed by the server pool per unit time.

8.1.3 Service Process Performance

Key performance measures of a service process can be expressed in terms of the usual flow rate, flow time, and inventory measures discussed earlier for general business processes:

1. **Flow rate–related measures of process capacity**
   - Throughput rate R is the average rate at which customers flow through the process, which is simply the smaller one of the customer arrival rate (demand) and the maximum processing rate (supply), so that

$$R = \min(R_i, R_p) \qquad \textbf{(Equation 8.2)}$$

   - Capacity utilization u is the average fraction of the server pool capacity that is busy processing customers and, as in Chapter 5, is given by

$$u = \frac{R}{R_p} \qquad \textbf{(Equation 8.3)}$$

Note that, if $R_p > R_i$ (i.e., the total processing capacity is sufficient to meet the demand for service), then the throughput rate is the same as the inflow rate ($R = R_i$) and the capacity utilization $u = R_i/R_p < 1$, so that some of the available capacity is unutilized. If, on the other hand, the inflow rate exceeds the processing rate, $R_i \geqslant R_p$, then $R = R_p$ and $u = 1$, so the resource pool is constantly busy processing customers. As we will see, $u < 1$ is essential for stability of a service process, if there is any variability in inflow and processing times at all.

- **Safety capacity** *R_s is the excess processing capacity (supply) available to process the customer arrivals (demand)*, and is given by

$$R_s = R_p - R \qquad \textbf{(Equation 8.4)}$$

Note that, if $R_p > R_i$ then $R = R_i$ and safety capacity $R_s = R_p - R_i > 0$. However, if $R_i > R_p$, then $R = R_p$ and $R_s = 0$, there is no safety capacity: All the available capacity is busy processing arrivals. As we will see, some safety capacity is essential to deal with any variability in arrival and processing times. The concept of safety capacity is the make-to-order process equivalent of safety inventory in make-to-stock processes. Both represent cushions that ensure availability of products and processes in the event of excess demand or shortfall in supply. Note that the two conditions

$$\text{capacity utilization } u = \frac{R_i}{R_p} < 1$$

and

$$\text{safety capacity } R_s = R_p - R_i > 0$$

are identical. Both mean $R_p > R_i$ so that, on average, the supply of service is more than sufficient to handle the demand for service.

2. **Flow time-related measures of customer delay**
   - Waiting time T_i is the average time that a customer spends in queue (the input buffer) before being served.
   - Service (or processing) time T_p is the average time required to process a customer. Recall that it is the theoretical flow time of a customer who does not have to wait for service.
   - Total flow time T is the average time that a customer spends waiting in queue and in service, so that

$$T = T_i + T_p \qquad \textbf{(Equation 8.5)}$$

   - Flow time efficiency e is the proportion of time that a customer spends being served rather than waiting in queue, computed as $e = T_p/T$. Since waiting is a non-value-adding activity, flow time efficiency measures the value-adding fraction of time that a customer spends in the process.
3. **Inventory-related measures of customer queues**
   - Queue length I_i is the average number of customers waiting for service; it is the average inventory of inflow units in the input buffer.
   - Number of customers in process I_p is the average in-process inventory. Since each customer is processed by one server, I_p is also the average number of servers that are busy at any given time processing customers.
   - Total number of customers in the system I is then the average total inventory within the process boundaries, which includes customers in queue and those being served, so that $I = I_i + I_p$.

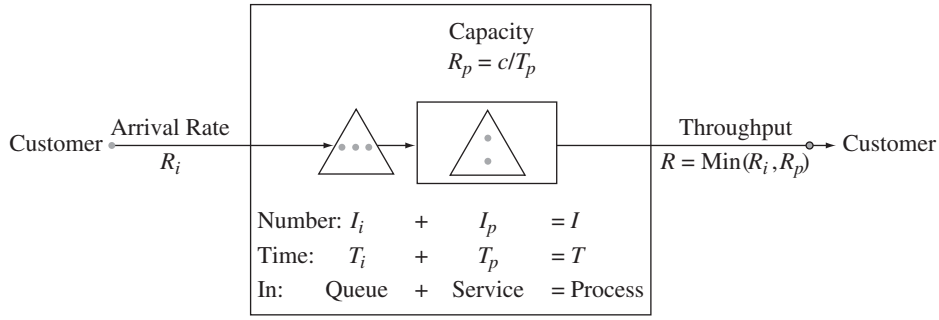


FIGURE 8.2 Service Process Flows, Delays, and Queues

Although we have described a service process attributes and its performance measures in more natural terms, we have already seen all of these concepts in the previous chapters, so we have kept the same general notation. In fact, most of the process flow concepts introduced in Chapters 3, 4, and 5 were inspired by the analysis of queuing systems, where Little's law was originally derived. For simplicity and consistency, we have employed a uniform notation to discuss a service process (or, more generally, a make-to-order process) as a special case of general business processes. This terminology and notation for make-to-order processes are summarized in Figure 8.2, which is very similar to Figure 6.1 in Chapter 6 for make-to-stock processes, the only difference being the absence of an output buffer since there is no inventory of completed orders, as they leave the process.

These concepts can be illustrated by Example 8.1 below, which is simply a continuation of Example 3.1 of Chapter 3.

EXAMPLE 8.1

Consider passenger arrivals at the Vancouver International Airport security checkpoint. Each passenger places his or her carry-on luggage on a moving belt to be scanned by an X-ray machine. So each passenger is a customer, and the X-ray scanner is the server, which processes customers in the FCFS order.

Suppose that, as before, a customer carries an average of 1.5 bags and that the X-ray machine can scan 18 bags per minute. So the X-ray machine can screen up to 12 passengers' bags per minute, and the average time to process a passenger is $T_p = 1/12$ minute, or 5 seconds. With one scanner, $c = 1$, the total processing rate is also $R_p = 12$ per minute. Prior to rescheduling flights, the peak arrival rate (from 9:10 a.m. to 9:30 a.m.) is estimated to be $R_i = 15$ per minute, so the average interarrival time is 1/15 minute, or 4 seconds. The passenger throughput rate is therefore $R = \min(15, 12) = 12$ per minute.

Since one customer arrives every 4 seconds and each customer requires 5 seconds to be processed, the queue will build up at the net rate of $R_i - R_p = 3$ per minute. Since $R_i > R_p$, the scanner capacity is insufficient to handle the passenger inflow. There is no safety capacity, ($R_s = 0$), and capacity utilization is 100% ($u = 1$), so the X-ray machine is constantly busy. Figure 8.3 shows passenger arrival and departure times and the time that each spends in the process, assuming that they arrive *exactly* 4 seconds apart and are processed in *exactly* 5 seconds. It is clear that each successive arrival will spend more and more time waiting and that the queue will keep building up. If this were to continue indefinitely, delays and queues would grow without limit, so in the long run, both T and I are infinite and we have an unstable process.

Customer number	Arrival time	Departure time	Time in process
1	0	5	5
2	4	10	6
3	8	15	7
4	12	20	8
5	16	25	9
6	20	30	10
7	24	35	11
8	28	40	12
9	32	45	13
10	36	50	14

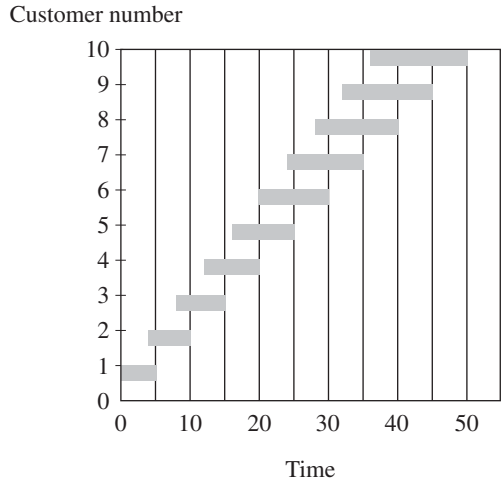


FIGURE 8.3 Flow Times with an Arrival Every Four Seconds

Recall that, after rescheduling flights, the customer arrival rate drops to $R_i = 10$ per minute (see Table 3.3), or one every 6 seconds. The new arrival rate is below the processing rate $R_p = 12$ per minute (or one every 5 seconds), and the scanner will be able to handle all arrivals. Now the throughput rate becomes $R = \min(10, 12) = 10$ per minute, the scanner has safety capacity of $R_s = R_p - R_i = 12 - 10 = 2$ per minute, and the capacity utilization is $u = R_i/R_p = 10/12$, or 0.8333. Thus, the scanner is busy 83.33% of the time and idle the remaining 16.67% of the time. Equivalently, since the interarrival time is 6 seconds and the processing time is 5 seconds, the scanner is busy 5 seconds out of every 6 seconds between arrivals, so it is busy 5/6 or 83.33%, of the time. Also note that 83.33% of the time there is one customer being scanned and that 16.67% of the time there is none, so the average number of customers being scanned at any instant is $I_p = (1)(0.833) + (0)(0.167) = 0.833$. Since each customer is being scanned by one server, $I_p = 0.833$ is also the average number of servers (out of one) busy processing at any instant. Finally, note that if the interarrival and processing times are *exactly* 6 and 5 seconds, respectively, no one has to wait in queue, so $I_i = 0$ and $T_i = 0$ as well (see Figure 8.4).

Customer number	Arrival time	Departure time	Time in process
1	0	5	5
2	6	11	5
3	12	17	5
4	18	23	5
5	24	29	5
6	30	35	5
7	36	41	5
8	42	47	5
9	48	53	5
10	54	59	5

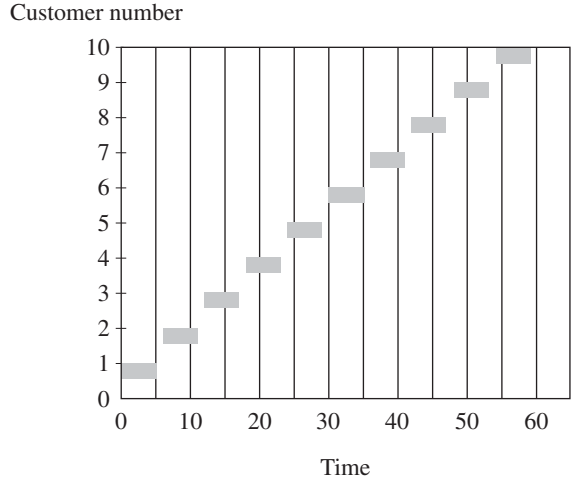


FIGURE 8.4 Flow Times with an Arrival Every Six Seconds

This example brings out an obvious but important reason why delays and queues occur: *If the interarrival and processing times are constant, queues will develop if and only if the arrival rate is greater than the processing rate.*

Although the arrival rate may exceed the processing rate over some time intervals, the long-run stability of a service process requires that, on average, the processing rate (supply of service) be sufficiently high to be able to process all arrivals (demand for service). This requirement will be called the **stability condition**, which states that *the average processing rate R_p be greater than the average arrival rate R_i* . Equivalently, it means the capacity utilization $u < 1$, or the safety capacity $R_s > 0$. Note that this terminology is consistent with discussion in Section 3.4 in Chapter 3, where we called a process stable if its average outflow rate equals the average inflow rate, which was then called the throughput rate.

The stability condition is necessary for limiting delays and queues; if it is not satisfied, delays and queues will grow without limit. We will assume throughout this chapter that the stability condition holds.

8.1.4 Relationships between Performance Measures

With the throughput rate $R = \min(R_i , R_p)$, we can apply Little's law to derive relationships between the number of customers and their flow times at various stages of the process. Thus, the average waiting time in queue (T_i) and the average number of customers in queue (I_i) relate as

$$I_i = R \times T_i \qquad \textbf{(Equation 8.6)}$$

and the average time in service (T_p) and the average number of customers in service (I_p) as

$$I_p = R \times T_p \qquad \textbf{(Equation 8.7)}$$

so the average total flow time ($T = T_i + T_p$) and the average total number of customers in the system ($I = I_i + I_p$) as

$$I = R \times T \qquad \textbf{(Equation 8.8)}$$

Recall that the capacity utilization is defined as $u = R/R_p$, where $R_p = c/T_p$, so that $u = RT_p/c$. By Little's law, $I_p = R \times T_p$, so we have

$$u = \frac{I_p}{c} \qquad \textbf{(Equation 8.9)}$$

Since I_p, the average number of customers in service, is also the average number of busy servers, Equation 8.9 provides an alternate and more intuitive definition of capacity utilization: It is the average fraction of the server pool that is busy processing customers. Equivalently, it is the average fraction of time that each server in the pool is busy.

As we saw in Example 8.1, after staggering flights, the capacity utilization drops from $u = 1$ to 0.8333, and the queue disappears. Thus, making the process stable by reducing its capacity utilization (or, equivalently, by increasing the safety capacity) improves the process performance in terms of reduced delays and queues. Note that the capacity utilization $u = R_iT_p/c$ can be decreased (or the safety capacity $R_s = c/T_p - R_i$ can be increased) by the following means:

- Decreasing the average inflow rate R_i
- Decreasing the average processing time T_p
- Increasing the number of servers c

We have already seen in the X-ray scanner example how staggering flights reduced the passenger arrival rate at the security check. The processing time could be reduced by purchasing a faster scanner, hiring better-trained security officers, restricting the number of carry-on bags per customer, and so forth. Finally, installing a second belt and scanner would double the processing capacity and reduce the capacity utilization by half.

Naturally, process managers prefer high capacity utilization and low idle time, because it means fuller utilization of the available resources. In fact, ideally, they would prefer 100 percent utilization with $R_p = R_i$, and there is no idle capacity! If the interarrival and processing times were constant, there still will be no queues, since the process will have (just) enough capacity to handle all arrivals. Thus, in Example 8.1, if the interarrival and processing times are *exactly* 4 seconds, the X-ray scanner will be able to handle up to $R_i = R_p = 15$ passenger arrivals per minute (one every 4 seconds). However, as we will see in the next section, if there is *any* variability in arrival and processing times at all, then it is essential that there is some safety (or unutilized) capacity ($R_s > 0$ or $u < 1$); otherwise, queues will grow indefinitely. In fact, as we will see in the next section, if there is any variability in the interarrival and/or processing times, queues will develop, *even if* there is excess capacity.

8.2 EFFECT OF VARIABILITY ON PROCESS PERFORMANCE

Our analysis of Example 8.1 assumed that interarrival times and processing times are known and constant. We saw that in the absence of variability, we can eliminate any waiting by ensuring that the processing rate exceeds the arrival rate—that is, by keeping some safety capacity. However, service (and other make-to-order) processes are characterized by a high degree of variability in customer arrival and processing times. For example, at a call center, customers do not call in at fixed intervals, nor does a CSR spend exactly the same amount of time with each caller. Similarly, at the airport security check, passengers do not arrive at evenly paced times, nor do they all need exactly the same amount of time to be checked out. Such *unpredictable or random variability that a service process experiences is called* **stochastic variability**, to be distinguished from more predictable changes over longer periods of time, including trend and seasonal variability. Example 8.2 illustrates how stochastic variability may lead to a queue buildup even in a stable process where the order processing rate is greater than the order arrival rate.

EXAMPLE 8.2

Suppose that after staggering flights at Vancouver International Airport, the average passenger arrival rate at the security checkpoint drops to 10 per minute, or one every 6 seconds, whereas the average X-ray scanning rate is 12 per minute, or one every 5 seconds. However, now suppose the actual arrival and processing times are not constant but variable. In particular, suppose we observe the interarrival times of 10 passengers to be 10, 10, 2, 10, 1, 3, 7, 9, and 2 seconds, which average to 6 seconds, as before. Similarly, suppose we clock their processing times to be 7, 1, 7, 2, 8, 7, 4, 8, 5, and 1 second, which average to 5 seconds, again as before. With these observations, let us track times of passenger arrivals and departures and the number of customers in the process (in queue as well as in service) and plot them as in Figure 8.5.

Now note that while Passengers 1, 2, 3, and 5 do not have to wait, all others do. Comparing with Example 8.1, where the interarrival and service times were *exactly* 6

Customer number	Arrival time	Processing time	Time in process
1	0	7	7
2	10	1	1
3	20	7	7
4	22	2	7
5	32	8	8
6	33	7	14
7	36	4	15
8	43	8	16
9	52	5	12
10	54	1	11

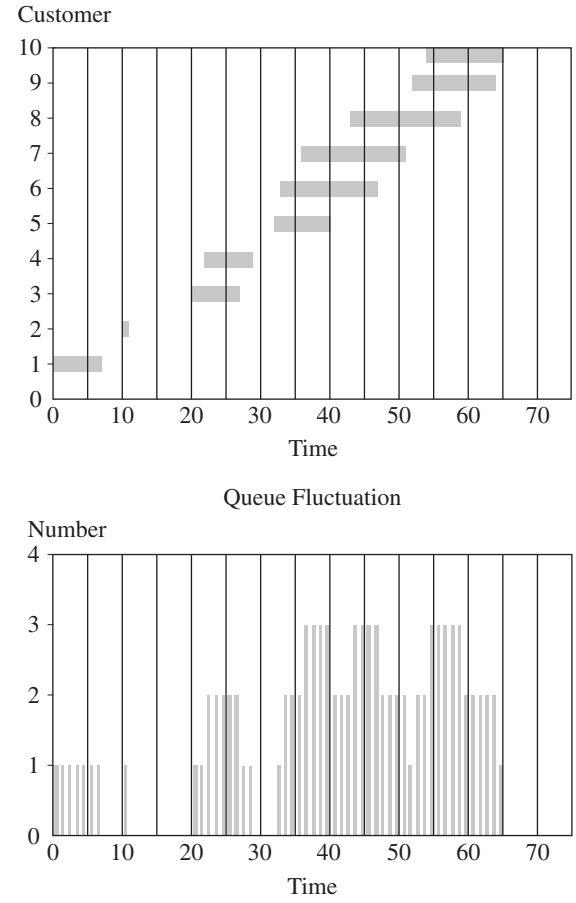


FIGURE 8.5 Effect of Variability in Arrivals and in Processing

and 5 seconds, respectively, and there was no waiting, we conclude that variability in these times leads to waiting and queues.

This example illustrates the second reason for delays and queues in service processes: Even under the stability condition (that the average processing rate is greater than the average arrival rate), variability in arrival and processing times may lead to customer delays and queues.

To understand the reason, note that Customer 2 needs only 1 second to be processed, and the next one does not come in for another 10 seconds, so the scanner is idle for 9 seconds after processing Customer 2. However, we cannot produce and store its scanning service in advance or store its idle capacity for later use to process Customer 4 when he arrives, who must therefore wait while Customer 3 is being processed. The basic problem is that service is nonstorable and processing capacity is perishable; if we do not use it, we lose it. In Example 8.2, the server is busy an average of 83.33% of the time, just as in Example 8.1. However, because of variability in arrival and processing times in Example 8.2, the server alternates between cycles of busy and idle periods, and its processing capacity during idle periods cannot be utilized to serve later customers who must therefore wait.

In general, with inflow variability, some customers have short interarrival times, while others have long interarrival times. Similarly, some customers have short processing times, while others have long processing times. When short interarrival times coin-

cide with long processing times, queues build up. In essence, this is due to an imbalance between inflows and outflows and the inability to shift processing times between customers and across time. The situation could be mitigated if we could match the arrival times (demand) and processing times (supply), leading to more uniform capacity utilization, as we see next.

Effect of Synchronization between Arrival and Processing Times Variability alone does not cause queues to build up. Queues build up because the variability in processing times is *independent* of the variability in interarrival times. In Example 8.2, Customer 4 has a short interarrival time and arrives 2 seconds after Customer 3. Customer 3, however, has a long processing time of 7 seconds, causing Customer 4 to wait for 5 seconds. If interarrival and processing times could be synchronized (or positively correlated), waiting times would be reduced significantly. Indeed, if short interarrival times of customers could be coupled with short processing times of their predecessors and long interarrival times are coupled with long processing times, queues will not build up, as the next example shows.

EXAMPLE 8.3

Suppose in Example 8.2 that the processing times of the 10 arrivals can be rearranged to be 8, 8, 2, 7, 1, 1, 7, 7, 4, and 5 (while keeping their arrival times at 0, 10, 20, 22, 32, 33, 36, 43, 52, and 54, as before). Now only Passenger 10 will have to wait for 3 seconds despite variability in interarrival and processing times. The first passenger leaves at time 8, while the second arrives at time 10. The second passenger leaves at time 18, while the third arrives at time 20, and so forth. Note that, as before, the capacity utilization factor is still 0.833. In fact, since the processing times were merely reshuffled from those in Example 8.3, they have the same mean and variability as before. However, better synchronization between supply and demand has led to significantly less waiting.

Unfortunately, because of the idiosyncratic nature of individual processing requirements in service and other make-to-order operations, we cannot interchange processing times across customer arrivals over time and achieve synchronization to reduce their waiting times. Later in Section 8.8, we will study strategies for achieving some degree of synchronization between interarrival and processing times.

For now, we may state the following qualitative observations about the corrupting influence of stochastic variability on process performance: Queues form when the customer arrival rate is—at least temporarily—greater than the rate at which customers are being processed. If the interarrival and/or processing times display any variability that is not synchronized, queues may form *even if* the average interarrival time is longer than the average processing time—that is, even when there is some safety capacity and the capacity utilization is less than 100%.

To summarize the key insights of this and the preceding sections, the main causes of delays and queues are the following:

1. High capacity utilization $u = R_i/R_p$ or low safety capacity $R_s = R_p - R_i$, which is due to
   - High arrival rate R_i
   - Low service rate $R_p = c/T_p$, which may be due to low c and/or high T_p
2. High, unsynchronized variability in
   - Interarrival times
   - Processing times

In Chapters 3 through 5, we learned how the process performance depends on the average flow times and average flow rates. The key lesson of this section is that variability in these factors also matters. Even if the process has sufficient capacity to handle inflows on average, variability in these factors will degrade the process performance. The effect of both of these factors is studied in a greater detail in the next section.

8.3 DRIVERS OF PROCESS PERFORMANCE

The two key drivers of process performance— capacity utilization and stochastic variability—are determined by two factors:

1. The mean and variability of interarrival times
2. The mean and variability of processing times

In practice, interarrival times can be measured by tracking either the times of customer arrivals or the total number of arrivals during a fixed time period. Likewise, processing times can be measured for different customers. The mean interarrival and processing times can then be estimated by computing the averages. Variability in the interarrival and processing times can be measured by their variances (or standard deviations), which indicate their dispersion around the means. Higher standard deviation means greater variability. However, standard deviation alone may not provide a complete picture of variability. For example, if the mean processing time is 2 minutes, the standard deviation of 1 minute represents significantly more variability than if the mean processing time was 10 minutes. Therefore, we should measure variability in time relative to its mean. One such measure is obtained by computing *the ratio of the standard deviation to the mean,* which is called the **coefficient of variation**. We denote the coefficients of variation of interarrival and processing times by C_i and C_p, respectively. The greater the coefficient of variation, the more variable the time is in relation to its mean. We are now ready to indicate how variability and capacity utilization jointly affect process performance.

8.3.1 The Queue Length Formula

The following *approximate* expression shows how the average queue length I_i depends on the coefficients of variation C_i and C_p of interarrival and processing times as well as the capacity utilization $u = R_i/R_p$ and the number of servers c (details may be found in Chapter 8 of Hopp and Spearman [2008]):

$$I_i = \frac{u^{\sqrt{2(c+1)}}}{1-u} \times \frac{C_i^2 + C_p^2}{2} \qquad \textbf{(Equation 8.10)}$$

This equation will be referred to as the **queue length formula**—*which shows how the average queue length depends on the capacity utilization, number of servers and variability in interarrival and processing times.* Note that the average queue length I_i is a product of two factors. The first factor

$$\frac{u^{\sqrt{2(c+1)}}}{1-u}$$

captures the capacity utilization effect, which shows that the queue length increases rapidly as the capacity utilization approaches 100% (as u increases to 1). As the processing capacity approaches the arrival rate (or, equivalently, as the safety capacity approaches zero), the average queue length (and hence waiting time) approaches infinity.

The second factor,

$$\frac{C_i^2 + C_p^2}{2}$$

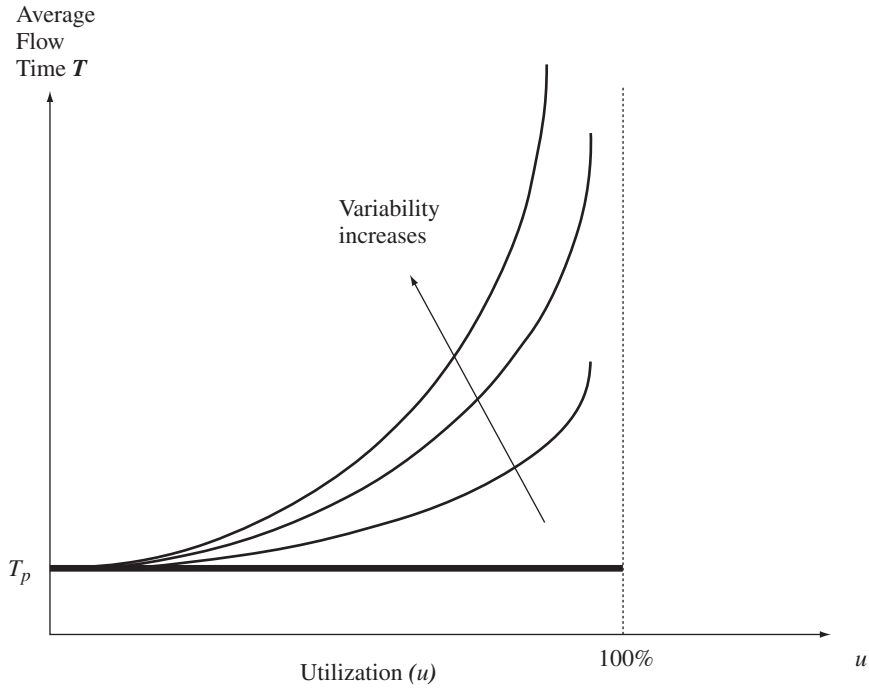


FIGURE 8.6 Throughput Delay Curve

captures the variability effect, which shows that the queue length increases as the variability in interarrival and processing times increases. Note that the effects of variability in interarrival and processing times on the queue length are similar and additive (which is due to the assumption that the interarrival and processing times are independent). The variability effect in the queue length formula shows that, whenever there is variability in arrivals or in processing—as is usually the case in practice—queues will build up and customers will have to wait, even if the processing capacity is, on average, sufficient to handle demand.

The queue length formula can be illustrated graphically in Figure 8.6 by the **throughput delay curve**, *which displays the average flow time as a function of capacity utilization.* It shows how the average flow time (waiting in queue and in process) increases with the capacity utilization for different levels of variability. In particular, it shows that the average flow time increases rapidly with capacity utilization and variability.

The queue length formula can be used to approximately compute the process performance measures, as illustrated in Example 8.4, which also illustrates the effect of increasing capacity on the process performance.

EXAMPLE 8.4

Suppose, as in Examples 8.2 and 8.3, that we observe the interarrival times of 10 passengers to be 10, 10, 2, 10, 1, 3, 7, 9, and 2 seconds. The average interarrival time is then 6 seconds (so the average arrival rate is $R_i = 1/6$ per second) with a standard deviation of 3.937 seconds, so its coefficient of variation is $C_i = 3.937/6 = 0.6562$. Similarly, if their processing times were observed to be 7, 1, 7, 2, 8, 7, 4, 8, 5, and 1 seconds, the average processing time can be computed to be 5 seconds (so the average processing rate is $R_p = 1/5$ per second) with a standard deviation of 2.8284 seconds, so the coefficient of variation is $C_p = 2.8284/5 = 0.5657$. Furthermore, since $R_i < R_p$, the throughput $R = R_i$.

With $c = 1$, and $u = R_i/R_p = 0.8333$, we can estimate the average number of passengers in queue by the queue length formula

$$I_i = \frac{0.8333^2}{1 - 0.8333} \times \frac{0.6562^2 + 0.5657^2}{2} = 1.5633$$

Thus, on average, there will be 1.5633 passengers waiting in line, even though, on average, we have sufficient processing capacity $R_p = 1/5$ per second to handle the inflow rate $R_i = 1/6$ per second, or we have safety capacity of $R_s = R_p - R_i = 1/5 - 1/6 = 0.0333$ per second.

To compute other performance measures, we can use Little's law and basic definitions. Thus, the average time that a passenger will spend in queue is $T_i = I_i/R_i = (1.5633)(6) = 9.38$ seconds. With $T_p = 5$ seconds in processing, on average, total time each passenger spends in the process is $T = 9.38 + 5 = 14.38$ seconds. Again, by Little's law, on average, total number of passengers in the process is $I = R \times T = 14.38/6 = 2.3967$. Equivalently, the average number of customers in process is $I_p = R \times T_p = 5/6 = 0.8333$, and $I = I_i + I_p = 1.5633 + 0.8333 = 2.3966$.

Suppose, in order to improve the process performance, we decide to increase the processing capacity by adding a second scanning machine. Now $c = 2$, the total processing rate doubles to $R_p = c/T_p = 2/5 = 0.4$ passengers per second, and capacity utilization is reduced to $u = R/R_p = 0.4167$ (so the safety capacity is increased to $R_s = R_p - R_i = 2/5 - 1/6 = 0.2333$ per second). Substituting these values in the queue length formula, the average number of passengers waiting in line becomes

$$I_i = \frac{0.4167^{\sqrt{2(2+1)}}}{1 - 0.4167} \times \frac{0.6562^2 + 0.5657^2}{2} = 0.075361$$

Other performance characteristics can be computed similarly. We summarize the results in the following table:

c	u	R_s	I_i	T_i	T	I
1	0.8333	0.0333	1.5633	9.38	14.38	2.3966
2	0.4167	0.2333	0.07536	0.45216	5.45216	0.9087

Thus, even with variability in arrivals and processing, reducing the capacity utilization (or increasing the safety capacity) improves the process performance in terms of reduced queues and delays.

The queue length formula identifies the two key drivers of process performance: capacity utilization and variability in interarrival and processing times. Along with Little's law, it permits us to compute various measures of process performance. However, it is important to remember that the queue length formula is only an approximation and not an exact relationship. To obtain an exact expression, one must make specific assumptions about probability distributions of the interarrival and processing times. The best-known and most tractable model for representing variability in these times is the exponential model that we outline next. It turns out that in this special case, and with one server, the queue length formula gives exact results.

8.3.2 The Exponential Model

In this model, the interarrival and processing times are assumed to be independently and exponentially distributed with means $1/R_i$ and T_p, respectively. Independence of interarrival and processing times means that the two types of variability are completely

unsynchronized. As for the exponential distribution, it is mathematically described in Appendix II of the text, while here we indicate key implications underlying this assumption. In essence, it represents complete randomness in interarrival and processing times. For example, if interarrival times are exponentially distributed with a mean of 6 seconds, then at any instant, the time until next arrival is completely independent of the time of the last arrival. This "memorylessness" property of arrivals holds in practice if there is an unlimited pool of potential arrivals, who make arrival decisions independently of one another over time. Similarly, if processing times are exponentially distributed with mean of 5 seconds, then regardless of how long a customer has already been processed, he should expect to spend yet another 5 seconds before being released; the remaining processing time is totally unpredictable on the basis of the past. Although the exponential interarrival time assumption is reasonable in practice, exponential processing time assumption is made mainly for analytical tractability.

If interarrival times are exponentially distributed with mean $1/R_i$, then the probability that the time between two arrivals will exceed any specific value t is given by $e^{-R_i t}$, where the mathematical constant $e = 2.718282$ is the base of the natural logarithm. This probability can be calculated by using the EXP function in Microsoft Excel, as shown in Example 8.5.

EXAMPLE 8.5

Suppose the time between consecutive passenger arrivals at the airport security X-ray scanner is exponentially distributed with mean of 6 seconds. The average arrival rate R_i is thus 1/6 per second (or 10 per minute). Then the probability that the time between two consecutive arrivals will exceed 10 seconds is given by $e^{-10/6} = \text{EXP}(-1.667) = 0.1888$. Similarly, if the time required to scan one customer's bags is exponentially distributed with mean of 5 seconds, then the likelihood that it will take no more than 3 seconds is given by $1 - e^{-3/5} = 1 - \text{EXP}(-0.6) = 0.451188$.

If the interarrival time is exponentially distributed with mean $1/R_i$, then the number of arrivals in any interval of duration t turns out to have Poisson distribution with mean $R_i t$. Again, Appendix II of the text describes the Poisson distribution. Intuitively, Poisson arrivals assumption models complete randomness: at any given time, regardless of the number and pattern of past arrivals, future arrivals in an interval of duration t will have Poisson distribution with mean $R_i t$.

It turns out that the exponential distribution assumption greatly facilitates mathematical analyses leading to exact formulas for computing process performance measures. For example, the standard deviation of the exponential distribution is the same as its mean, so its coefficient of variation is 1. If there is a single server, as in the X-ray scanner illustration, substituting $c = 1$ and $C_i = C_p = 1$ into the queue length formula yields

$$I_i = \frac{u^2}{1 - u}$$

from which all other performance measures can be calculated by using Little's law and the basic definitions summarized in Figure 8.2. In particular, the average total time that a customer spends in the process turns out to be

$$T = \frac{1}{R_p - R_i} = \frac{1}{R_s} \qquad \textbf{(Equation 8.11)}$$

That is, a customer's average flow time is inversely proportional to the safety capacity, so increasing safety capacity decreases average flow time.

With multiple servers ($c \geqslant 2$), exact formulas for computing the queue length and wait are also available but complicated, even with the exponential distribution assumption. These formulas are provided in Appendix for reference and are programmed in a Microsoft Excel spreadsheet called **Performance.xls**, which was developed by Professor John O. McCain of Cornell University and can be downloaded from the Prentice Hall's Web site at www.prenhall.com/anupindi. In fact, the spreadsheet computes performance characteristics of two types of processes: those with (1) finite and (2) infinite input buffer capacities. It requires specification of four inputs: the number of servers (c), the average arrival rate (R_i), the average processing rate of each server ($1/T_p$) and the buffer capacity (K). We will study implications of limited buffer capacity in Section 8.5, but for now, we can use the spreadsheet to perform the computations assuming infinite buffer capacity. It then calculates key performance characteristics such as the average number of customers waiting in queue and the average waiting time of each customer. We illustrate these calculations for the airport security example, this time assuming that the interarrival and processing times are exponentially distributed.

EXAMPLE 8.6

Recall that the average passenger arrival rate is $R_i = 10$ per minute and the average processing time of each is $T_p = 5$ seconds, so the average processing rate of each scanner is $1/T_p = 1/5$ per second or 12 per minute. With one scanner, the number of servers $c = 1$, so that the total processing rate $R_p = c/T_p = 12$ per minute. As before, the capacity utilization $u = R_i/R_p = 10/12 = 0.8333$ and the safety capacity $R_s = R_p - R_i = 12 - 10 = 2$/min. To obtain other performance measures, we use "Infinite Queue" worksheet and enter $c = 1$, $R_i = 10$, and $1/T_p = 12$. The spreadsheet yields $I_i = 4.167$, $T_i = 0.4167$ minute = 25 seconds, $I = 5$, and $T = 0.5$ minute = 30 seconds.

If we add another X-ray machine to improve the performance, we simply change $c = 2$, and the spreadsheet calculates $I_i = 0.175$, $T_i = 0.018$ minute = 1.08 seconds, $T = 0.101$ minute = 6.06 seconds, and $I = 1.008$, and $u = 0.4167$. The following table summarizes these results:

c	u	R_s	I_i	T_i	T	I
1	0.8333	0.0333	4.167	25	30	5
2	0.4167	0.2333	0.175	1.08	6.06	1.008

Now, we could ask if the reduction in a passenger's average waiting time from 25 seconds to 1 second is worth the extra cost of purchasing the second X-ray machine. Equivalently, the additional processing capacity has reduced congestion from five passengers down to one passenger. In the next section we will study economic tradeoffs involved in making these decisions. For now we emphasize that even though the process is stable ($u < 1$) and we do have safety capacity ($R_s > 0$), variability in arrival and processing times will result in customer delays and queues.

Note that, although qualitatively similar, the exact numerical results in Examples 8.4 and 8.6 are different. That is because in Example 8.4 we permitted arbitrary probability distributions of interarrival and processing times and used the queue length formula to obtain *approximate* results, whereas in Example 8.6 we obtained *exact* results but had to assume exponential distribution for interarrival and processing times. We conclude this section by emphasizing that the queue length formula is exact only for the exponential model and that too with one server.

8.4 PROCESS CAPACITY DECISIONS

As we saw in the previous section, increasing the processing capacity improves process performance in terms of reduced waiting times and queues. The inconvenience of waiting and the resulting customer dissatisfaction has financial implications in terms of lost reputation and future revenue if disgruntled customers decide to take their business elsewhere. Moreover, trucks or ships waiting to be loaded or unloaded in docks or harbor result in real economic costs of carrying the pipeline inventory. In this section, we consider the problem of determining optimal capacity that minimizes the total cost of providing service and the cost of waiting. Optimal capacity investment should balance the cost of additional capacity against the benefit of greater customer satisfaction. We illustrate by analyzing the staffing problem at the L. L. Bean call center.

EXAMPLE 8.7

Suppose the call center is currently staffed by only one CSR who takes an average of $T_p = 2.5$ minutes to process a call, so her processing rate is $1/T_p = 0.4$ customer per minute or 24 per hour. With number of servers $c = 1$, the processing capacity is also $R_p = 24$ per hour. Suppose customer calls come in at an average of 3 minute intervals, so the average arrival rate is $R_i = 1/3$ per minute or 20 per hour. Since $R_p > R_i$, the process is stable. However, variability in arrivals and processing will result in delays and queues. We will assume the interarrival and processing times are exponentially distributed so we can use the spreadsheet **Performance.xls** to calculate process performance characteristics.

Suppose L. L. Bean estimates that the retailer loses \$2 in sales for every minute that a customer has to wait on line for a CSR, in terms of dissatisfaction with service as well as the resulting impact on future sales to disgruntled customers. Since I_i is the average number of callers waiting in line, the average total cost of waiting will be $\$2I_i$ per minute or $\$120I_i$ per hour. Alternately, each caller waits an average of T_i minutes for a CSR at a cost of $\$2T_i$, while an average of R_i customers call in every minute, so the total cost of waiting is $\$2R_iT_i$ per minute or $\$120R_iT_i$ per hour. However, by Little's law, we know $R_iT_i = I_i$, so the waiting cost is again $\$2I_i$ per minute or $\$120I_i$ per hour. We can use **Performance.xls** spreadsheet to calculate I_i for different numbers of servers $c = 1, 2, \ldots$. Obviously, as we hire more servers, the cost of waiting will go down, while the cost of providing service will go up. Suppose each CSR is paid \$20 an hour, so with c servers, the hourly cost of providing service will be $\$20c$. The manager of the L. L. Bean call center would like to determine the optimal number of CSRs to minimize

$$\text{Total hourly cost} = \$20c + \$120I_i$$

We compute and tabulate this cost by using **Performance.xls** with following inputs:

$c = 1, 2, \ldots$.
$R_i = 20$/hour
$1/T_p = 24$/hour

And obtain hourly costs as summarized in the following table.

c	I_i	$\$20c$	$\$120I_i$	Total Hourly Cost
1	4.167	\$20	\$500.04	\$520.04
2	0.175	\$40	\$21.00	\$61.00
3	0.022	\$60	\$2.64	\$62.64
4	0.003	\$80	\$0.36	\$80.36

Thus, the total hourly cost of waiting and providing service is minimized when the number of CSRs is $c = 2$.

Alternately, L. L. Bean may be concerned with the *total* turnaround time T that a customer spends for the entire transaction, including waiting for a CSR *and* being served. In that case, L. L. Bean's problem is to determine c that minimizes

$$\text{Total hourly cost} = \$20c + \$120I.$$

The results are summarized below.

c	I	$20c$	$120I$	Total Cost
1	5.000	\$20	\$600.00	\$620.00
2	1.008	\$40	\$120.96	\$160.96
3	0.856	\$60	\$102.72	\$162.72
4	0.836	\$80	\$100.32	\$180.32

Again, $c = 2$ minimizes the total hourly cost of providing service and customer's total time spent in the system.

8.5 BUFFER CAPACITY, BLOCKING, AND ABANDONMENT

Thus far we have assumed that all arrivals get in and are eventually processed, as in the security checkpoint at the Vancouver International airport, so the throughput rate of the process is limited only by its inflow rate and the maximum processing rate. In this section, we consider situations in which some of the arrivals may not be able to enter the process at all, while some who do enter may choose to leave because of long delays before being served. We will then evaluate the process performance in terms of the throughput rate, waiting time, and queue length.

In many applications, there may be a *limit on the number of customers that can wait before being served*, which is called the **buffer (or waiting room) capacity**, to be denoted as K. When the buffer is full, *any new arrivals are turned away, which is called* **blocking**. For example, waiting space in a restaurant, barber shop, or the drive-in facility at a bank, storage bins for purchased parts, or telephone lines at a call center all have limited buffer capacity to accommodate customers waiting for service. In the L. L. Bean call center if there are two CSRs and six telephone lines, then at most four callers can be put on hold, so the buffer capacity is $K = 4$. Once the buffer is full, any new caller will get a busy signal and cannot enter the system. These blocked arrivals represent loss of business if they do not call back.

Moreover, even if they are able to join the queue, some of the *customers who have to wait long for service may get impatient and leave the process before being served,* which is called **abandonment**. Again, if they do not return, it means lost business.

To analyze these situations, we need to introduce some additional notation. *The average fraction of arrivals blocked from entering the process because the input buffer is full* is referred to as the **proportion blocked** (or **probability of blocking**) and is denoted by P_b. Thus, even though the potential customer arrival rate is R_i, only a fraction $(1 - P_b)$ gets in, so the net rate at which customers join the queue is $R_i(1 - P_b)$. Moreover, out of those customers who do get in, *a certain fraction P_a may abandon the queue, which is referred to as the* **proportion abandoning**, denoted as P_a. Thus, the net rate at which customers actually enter, get served, and exit the process is $R_i(1 - P_b)(1 - P_a)$, and the resulting throughput rate can then be calculated as

$$R = \min[R_i(1 - P_b)(1 - P_a), R_p] \qquad \textbf{(Equation 8.12)}$$

Thus, fractions of customers blocked and abandoned are important measures of process performance because they affect the throughput rate which in turn impacts financial measures of process performance. With blocking and abandonment, T_i now refers to the waiting time of only those customers who get into the system and are served. Note that with limited buffer capacity, regardless of the magnitudes of inflow and processing rates, the queue will never exceed K, thus assuring that the process will be stable.

8.5.1 Effect of Buffer Capacity on Process Performance

With finite buffer capacity, but without abandonment ($P_a = 0$), the Finite Queue worksheet of **Performance.xls** can be used to calculate various performance measures for given values of the number of servers c, buffer capacity K, arrival rate R_i, and the processing rate of each server $1/T_p$. Specifically, the spreadsheet calculates the probability of blocking P_b, the average number of customers in queue I_i and in the system I, the average waiting time of a customer in queue T_i and in the system T, the capacity utilization u, and so forth. We illustrate these computations for the call center application.

EXAMPLE 8.8

Suppose that the L. L. Bean's call center is currently staffed by one CSR who takes an average of 2.5 minutes to handle a call and suppose that calls come in at an average rate of 20 per hour. Furthermore, suppose there are five telephone lines, so that, at most, four customers can be put on hold. L. L. Bean would like to estimate the proportion of callers who will get a busy signal and are thus lost to the competition. They would also like to know the average time that a customer has to wait for a CSR to become available. Finally, they would like to know the effect of adding more telephone lines on various performance measures.

In this case, we have a service process with finite buffer capacity, and we are given the following information:

Number of servers $c = 1$

Buffer capacity $K = 4$

Arrival rate $R_i = 20$ per hour

Processing time $T_p = 2.5$ minutes or the processing rate of each server $1/T_p = 1/2.5 = 0.4$ per minute or 24 per hour. With this data input into the Finite Queue worksheet of **Performance.xls** spreadsheet, we get the following measures of performance:

Probability of blocking $P_b = 10.07\%$

Average number of calls on hold $I_i = 1.23$

Average waiting time of a caller on hold $T_i = 0.06835$ hours $= 4.1$ minutes

Average total time that a caller spends in the system $T = T_i + T_p = 4.1 + 2.5 = 6.6$ minutes

Average total number of customers in the system $I = 1.98$

Thus, on average, about 10% of all callers will get a busy signal and go elsewhere, and about 90% get through. If no one abandons the queue ($P_a = 0$), the throughput rate will be

$$R = \min[R_i(1 - P_b), R_p] = \min[20 \times (1 - 0.1007), 24] = 17.99 \text{ calls/hour}$$

and the average server utilization will be

$$u = \frac{R}{R_p} = 17.99/24 = 0.7495$$

Table 8.1 Effect of Buffer Capacity on Process Performance

Number of telephone lines n	5	6	7	8	9	10
Number of servers c	1	1	1	1	1	1
Buffer capacity $K = n - c$	4	5	6	7	8	9
Blocking probability P_b (%)	10.07	7.74	6.06	4.81	3.85	3.11
Throughput R (units/hour)	17.99	18.46	18.79	19.04	19.23	19.38
Average number of calls in queue l_i	1.23	1.52	1.79	2.04	2.27	2.48
Average wait in queue T_i (minutes)	4.10	4.95	5.73	6.44	7.09	7.68
Capacity utilization u	0.75	0.77	0.78	0.79	0.80	0.81

Thus, the CSR is busy only about 75% of the time and idle for about 25% of the time. Because of variability, however, there will be an average of 1.23 callers on hold and 10% of all callers (or two per hour) will get a busy signal, resulting in lost sales.

To study the effect of adding more telephone lines, we simply change the value of K and see how it affects key performance measures. Table 8.1 summarizes the results (rounded up to two decimal places).

Note that, as the buffer capacity (number of telephone lines) is increased, the blocking probability declines, and more callers are able to get into the system. Interestingly, however, the average waiting time of the callers who do get in increases. Thus, increasing the buffer capacity has two opposing effects: increasing the process throughput but also increasing the average waiting time of customers served. The optimal buffer size should take into account the financial impact of both, as we study in the next subsection.

8.5.2 The Buffer Capacity Decision

Customers blocked from entering the call center cost the retailer potential revenue if they do not call back. If they do enter they may have to wait on hold, during which the call center may have to pay telephone charges. Long waits also mean customer dissatisfaction and some customers abandoning the queue, again resulting in lost potential revenue. Thus, each of the operational performance measures, that is, blocking, abandonment, queues, and delays has a direct bearing on economic measures, which are, revenues and costs. Capacity investment decisions should balance them all. We illustrate with L. L. Bean's problem of choosing the number of telephone lines to lease.

EXAMPLE 8.9

Continuing Example 8.8, suppose that any caller who receives a busy signal hangs up and orders from a competitor. L. L. Bean estimates the average cost of lost sales to be $100 per customer.

Furthermore, suppose that after a customer call gets in, each minute spent waiting on hold costs the retailer $2 in terms of lost goodwill (which may affect future sales). If leasing each telephone line costs $5 per hour, how many lines should the call center lease?

Note that the call center incurs four types of costs:

1. Cost of the CSR's wages, say, $20 per hour
2. Cost of leasing a telephone line, assumed to be $5 per line per hour
3. Cost of lost contribution margin for callers getting busy signals, assumed to be $100 per blocked call
4. Cost of waiting by callers on hold, assumed to be $2 per minute per customer

Table 8.2 Effect of Buffer Capacity on Total Operating Cost

Number of telephone lines n	5	6	7	8	9	10
Number of CSRs c	1	1	1	1	1	1
Buffer capacity $K = n - c$	4	5	6	7	8	9
Cost of CSR's wages (\$/hour) = $20c$	20	20	20	20	20	20
Cost of telephone lines (\$/hour) = $5n$	25	30	35	40	45	50
Probability of blocking P_b (%)	10.07	7.74	6.06	4.81	3.85	3.11
Margin lost due to blocking (\$/hour) = $100\ R_i\ P_b$	201.4	154.8	121.2	96.2	77.0	62.2
Average number of calls waiting on hold I_i	1.23	1.52	1.79	2.04	2.27	2.48
Average cost of waiting (\$/hour) = $120I_i$	147.6	182.4	214.8	244.8	272.4	296.4
Total cost of service, blocking, and waiting (\$/hour)	394.0	387.2	391.0	401	414.4	428.6

Now, with one CSR and five telephone lines, $c = 1$ and $K = 4$, the cost of the server is $\$20c = \$20$ per hour, and the cost of leasing telephone lines is $\$5(K + c) = (5)(4 + 1) = \$25$ per hour.

We determined in Example 8.8 that the average number of customers blocked because of busy signals is

$$R_i\ P_b = (20)(0.1007) = 2.014/\text{hour}$$

The contribution margin lost because of blocking is therefore

$$\$100\ R_i\ P_b = (100)(2.014) = \$201.40/\text{hour}$$

Performance.xls gave the average number of customers on hold as $I_i = 1.23$. If each waiting customer costs \$2 per minute, or \$120 per hour, the hourly waiting cost will be

$$\$120I_i = (120)(1.23) = 147.6/\text{hour}$$

The total operating cost, therefore, is

$$\$(20 + 25 + 201.4 + 147.6) = \$394/\text{hour}$$

Increasing the number of telephone lines increases the buffer capacity K, and, as above, we can compute the total cost per hour, as summarized in Table 8.2.

Thus the total cost is minimized when the number of telephone lines is $n = 6$ or the optimal buffer capacity is $K = 5$. Leasing one more line not only costs more but also increases the cost of waiting time experienced by callers who do get in. In this instance, the waiting time of a caller is so expensive that the firm is better off not serving some customers at all than first admitting them and then subjecting them to long waits. Conversely, leasing one fewer line is also nonoptimal because it leads to more blocking and a greater loss of contribution margin on customers that are turned away than the saving in the cost of leasing the telephone line or customer waiting. The optimal buffer size thus correctly balances these costs.

It is interesting to note that, although limiting the buffer capacity denies access to new arrivals when the buffer is full, these customers would have had to wait long if they were allowed to get in, so they may be better off not getting in at all! An approach to imposing buffer limitation would be to inform callers that their wait may be long and hence they should call back later or, better yet, the service provider will call them back later. That would improve service to customers who do get in, without affecting those who are blocked.

8.5.3 Joint Processing Capacity and Buffer Capacity Decisions

In section 8.4, we determined optimal processing capacity c assuming unlimited buffer. In section 8.5.2, we determined optimal buffer capacity K for a given processing capacity c. In this section we determine both the processing capacity c and the buffer capacity K to minimize the total cost, which consists of the cost of servers and the buffer capacity as well as the loss due to customer blocking and waiting. We illustrate by determining the optimal number of CSRs and telephone lines to install in the call center example.

EXAMPLE 8.10

As before, suppose the call center has an average of 20 incoming calls per hour. A caller who gets a busy signal is blocked for an opportunity loss of \$100, and each minute spent by a customer on hold costs \$2 in terms of lost goodwill. Recall that each CSR takes 2.5 minutes to process one call and is paid \$20 per hour. Suppose leasing a telephone line costs \$5 an hour. The problem is to determine the optimal number of CSRs and telephone lines.

The total hourly cost, then, consists of the following:

- Cost of CSR's wages: \$20c
- Cost of line charges: \$5$(K + c)$
- Cost of lost sales due to blocking: \$100$R_i P_b$
- Cost of waiting: \$120I_i

The problem is to determine c and K that minimizes

$$\text{Total hourly cost} = \$20c + \$5(K + c) + \$100R_i P_b + \$120I_i$$

With $R_i = 20$/hour, $1/T_p = 24$/hour and different values of c and K, the spreadsheet **Performance.xls** provides values of P_b and I_i, as summarized in Tables 8.3 and 8.4. Substituting them in the total hourly cost formula above yields Table 8.5.

Table 8.3 Effect of Buffer and Processing Capacity on the Blocking Probability

P_b	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
$c = 1$	27.47%	18.63%	13.44%	10.07%	7.74%	6.06%
$c = 2$	6.22%	2.53%	1.04%	0.43%	0.18%	0.07%
$c = 3$	1.16%	0.32%	0.09%	0.02%	0.01%	0.00%
$c = 4$	0.18%	0.04%	0.01%	0.00%	0.00%	0.00%
$c = 5$	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 8.4 Effect of Buffer and Processing Capacity on the Waiting Line

I_i	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
$c = 1$	0.27	0.60	0.92	1.23	1.52	1.79
$c = 2$	0.06	0.11	0.14	0.16	0.17	0.17
$c = 3$	0.01	0.02	0.02	0.02	0.02	0.02
$c = 4$	0.00	0.00	0.00	0.00	0.00	0.00
$c = 5$	0.00	0.00	0.00	0.00	0.00	0.00

Table 8.5 Effect of Buffer and Processing Capacity on the Total Hourly Cost

Total Cost	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
$c = 1$	$612.42	$479.11	$419.06	$394.00	$387.20	$391.00
$c = 2$	$186.91	$123.88	$102.79	$ 97.60	$ 98.60	$102.02
$c = 3$	$104.65	$ 93.60	$ 94.27	$ 98.10	$102.78	$107.70
$c = 4$	$108.86	$111.07	$115.50	$120.38	$125.35	$130.35
$c = 5$	$130.51	$135.12	$140.05	$145.04	$150.04	$155.04

It follows the lowest total hourly cost of $93.60 is attained at $c = 3$ and $K = 2$, so L. L. Bean should hire three CSRs and lease five telephone lines.

In this economic analysis, we have assumed specific values for the cost of lost sales due to blocking and the cost of customer's waiting time in queue. In practice, these costs are usually difficult to estimate. For example, how can one place a dollar value on the physical and mental pain suffered by a patient waiting in a hospital? Even at the business level, it is difficult to estimate future sales lost due to a customer dissatisfied with long waits who in turn may share his experience with friends. In such situations, instead of estimating and minimizing costs, the process manager could choose to set limits on the proportion of arrivals blocked and the average waiting time of a customer as policy variables and look for a combination of the buffer and processing capacities that would provide acceptable values of these performance measures.

8.6 PERFORMANCE VARIABILITY AND PROMISE

Our entire discussion of performance measurement and improvement has focused on the average queue length and the average waiting time. In this section, we study why considering only the average values of these performance measures may not be sufficient.

In a service process, the average waiting time includes both customers with very long waits and those with short or no waits. Now a customer who in fact had to wait 30 minutes for service is not likely to be comforted to know that the average waiting time of all customers was only ten minutes, and in fact 20 percent of all customers did not have to wait at all! Typically, customers' tolerance for waiting decreases with the duration of the wait. Those who have to wait for longer times are disproportionately more dissatisfied than those who have to wait for shorter times. Ideally, we would like to look at the entire probability distribution of the waiting time across all customers, not just its average. At least, we would like to know the probability that a customer may have to wait longer than a specified duration that we consider acceptable. It is important to know, therefore, what fraction of customers may experience an extraordinarily long waits, because that would highlight problematic or unacceptable service. Thus, we need to focus on the *upper tail* of the probability distribution of the waiting time, not just its average value.

In a single-phase single server service process and exponentially distributed interarrival and processing times, it turns out that the actual total time that a customer spends in the process is also exponentially distributed with mean T. Therefore, as in Appendix II,

$$\text{Prob(Total time in process} > t) = e^{-t/T} = \text{EXP}(-t/T)$$

is the proportion of arrivals who will have to wait for more than t time units. To illustrate, consider the following example.

EXAMPLE 8.11

Suppose Walgreen drug store's pharmacy is staffed by one pharmacist, Dave. Suppose on average $R_i = 20$ customers come to the pharmacy every hour to get their prescriptions filled. Suppose Dave takes an average of $T_p = 2.5$ minutes to fill a prescription, so his processing rate is $R_p = 24$ customers per hour. If we assume exponentially distributed interarrival and processing times, we have a single-phase, single-server exponential model of Section 8.3.2, so by Equation 8.11, the average total time in the process is given by $T = 1 / (R_p - R_i) = 1/ (24 - 20) = 0.25$ hours or 15 minutes. The fraction of customers who will spend more than $t = 15$ minutes is EXP $(- t/T)$ = EXP$(-1) = 0.3679$, so about 37 percent of customers will need more than 15 minutes to have their prescriptions filled. More seriously, the fraction of customers who will spend more than $t = 45$ minutes will be EXP$(- 45/15)$ = EXP$(-3) = 0.0498$, so about 5 percent of customers will spend more than 45 minutes in the drug store! For these 5 percent of customers, it is cold comfort to know that an average customer spends only 15 minutes in the system. These 5 percent of customers will also be the ones who will complain most bitterly about the delay and affect Walgreen's business. Thus, in addition to the average value, the 95th percentile of the distribution of the actual total time in process provides important information about the process performance. It qualifies the average total time of 15 minutes with a caution that 5 percent of customers will experience delays of 45 minutes or more.

Since a large fraction of customers may experience delays longer than the average T, the process manager may not wish to announce T as what most customers will experience (even though that is what a typical customer is expected to spend in the process). As a service promise, we may wish to quote that customers will be served within some conservatively high duration, say T_d, so that we will meet that promise most of the time. This **promised duration** T_d is the *time period within which the product or service will be delivered with a high probability.* For example, if Dave promises to fill prescriptions within half an hour, we have $T_d = 30$, even though average is $T = 15$ minutes. In that case, the proportion of customers who will have to wait more than the promised time is

$$\text{EXP}(-T_d/T) = \text{EXP}(-30/15) = 0.1353.$$

Thus, Dave will not be able to keep his promise on 13.53% of customers, that is, he will deliver on promise 86.47% of the time.

Analogous to Chapter 7, we may define the proportion of customers that will be served during the promised duration T_d as customer service level SL. Thus, by promising a 30 minute turnaround time, Dave is providing 86.47% customer service.

Conversely, we may set a service level (SL), and derive the corresponding due date T_d to promise such that

$$\text{Prob(Total time in process} \leq T_d) = SL \qquad \textbf{(Equation 8.13)}$$

or

$$\text{EXP}(-T_d/T) = 1 - SL$$

or

$$T_d = -T\, ln\, (1 - SL)$$

where ln denotes the natural logarithm with $e^{\ln(x)} = x$.

We can then define safety time (analogous to safety inventory and safety capacity) as

$$T_s = T_d - T \qquad \textbf{(Equation 8.14)}$$

or

$$T_d = T + T_s$$

Thus, **safety time** *is the time margin that we should allow over and above the expected time to deliver service in order to ensure that we will be able to meet the promised duration with high probability.* Clearly, the larger the safety time, the greater the probability of fulfilling the promise.

EXAMPLE 8.12

Suppose Dave wishes to promise time T_d such that he will be able to keep his promise on 90 percent of the customers. Thus he is choosing $SL = 0.90$, when the average time required is $T = 15$ minutes. The necessary due date will be

$$T_d = -(15)\, ln(0.10) = 34.54 \text{ minutes}.$$

Therefore, Dave should announce that 90 percent of his customers will get their prescriptions filled within 35 minutes of arrival. Even though he expects to fill prescriptions within 15 minutes of submission on average, he is allowing an additional 20 minutes of safety time to ensure serving 90 percent of his customers within the promised time.

8.7 CUSTOMER POOLING AND SEGREGATION

As we have seen in preceding sections, investment in buffer capacity and safety capacity improves process performance. In this section, we will learn how performance can also be improved by pooling arrivals in some cases and segregating them in others, coupled with the necessary resource flexibility and specialization.

8.7.1 Pooling Arrivals with Flexible Resources

Given the number of servers, we may choose to organize them so arrivals form separate lines, each processed by a dedicated server. Alternately, we can have them join one single line that is served by the entire pool of servers. Given all else the same, consolidating lines into one improves the process performance. To illustrate, reconsider the airport security example.

EXAMPLE 8.13

Suppose the airport authority has decided to invest in a second X-ray scanner in an effort to reduce passenger waiting times. Now the operations manager must decide how best to utilize the two scanners. Specifically, she has two choices of process design, shown in Figure 8.7:

- In Design A, customer arrivals are evenly and randomly split into two streams, one for each X-ray scanner, so each scanner has its own queue. (This could be done, for example, by assigning one scanner to check out passengers on one half of the flights and the other for the remaining half. Alternately, the two scanners could be physically separated at two ends of the terminal, so that on average half the passengers choose one and the other half chooses the other and once chosen they cannot switch.)
- In Design B, all passengers join a single queue, and each passenger is scanned by the next available agent.

Assuming that both designs require the same resource investment, which one would yield better process performance in terms of flow time and resource utilization?

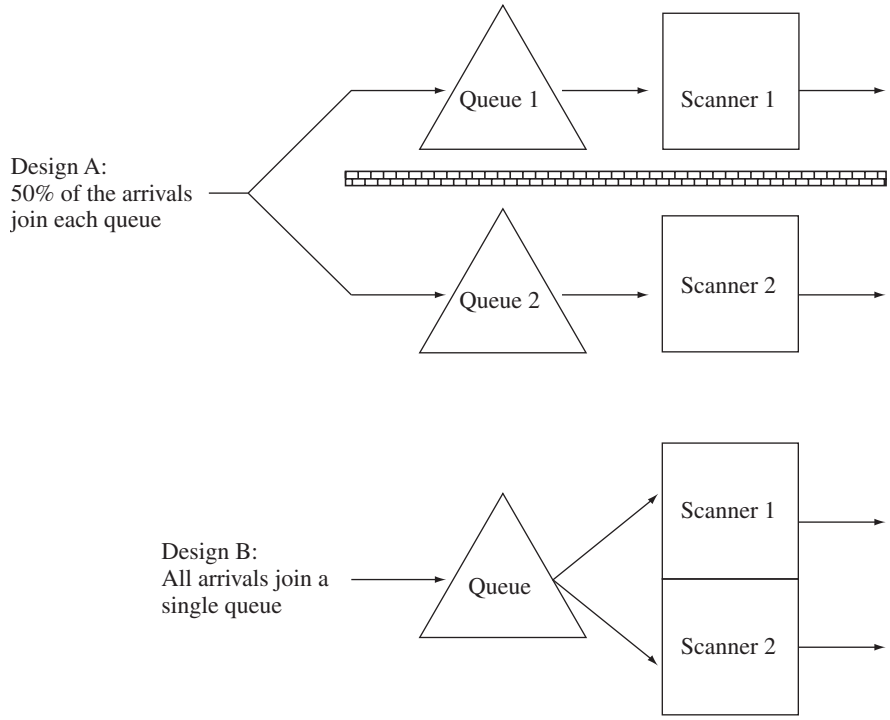


FIGURE 8.7 Pooling Capacity

With Process Design A, we have two independent and identical single server processes, each with one-half the arrival rate as before. Thus, each process has the arrival rate $R_i = 10/2 = 5$ per minute, while the average processing time of each is $T_p =$ 5 seconds (or the processing rate is $R_p = 1/5$ per second, or 12 per minute), as before. We may use Equation 8.11 to determine the total flow time T or **Performance.xls** spreadsheet with $R_i = 5$ per minute, $1/T_p = 12$ per minute, and $c = 1$ as inputs. The spreadsheet calculates various performance characteristics, including total flow time $T = 0.1429$ minute, or 8.57 seconds. This is a significant improvement over $T = 30$ seconds that we saw in Example 8.4 with one scanner handling all arrivals. Note that capacity utilization of each process now reduces by 50% to 0.4167, so each scanner is busy only 41.67% of the time.

Now consider Process Design B, where all customers join a single queue at the arrival rate $R_i = 10$ per minute and are served by the next available one of the two scanners in $T_p = 5$ seconds. . In the **Performance.xls** spreadsheet, we enter $c = 2$, $R_i = 10$ per minute, and $1/T_p = 12$ per minute, and get $T_i = 0.0175$ minute, or 1.05 seconds, and $T =$ 6.06 seconds, which represents a 33% improvement over design A. Other measures are also significantly better than with Design A. Note, however, that the capacity utilization is $u = R_i/R_p = 10/24 = 0.4167$, which is the same as for Design A, so each scanner is again busy for 41.67% of time (which is also the average fraction of scanners that are busy).

This example illustrates advantages of **pooling capacity**—*the sharing of available capacity among various sources of demand (or arrivals)*—in terms of reduced waiting times and queues. With two separate queues, it is possible that one server has customers waiting, while the other server is idle. With pooling the queues, a server will be idle only if there is no one in the system, so we are making a fuller use of the available capacity. Consider, too, the fact that in Design A, the waiting time of a customer is dependent on the pro-

cessing times of those ahead in the queue. When there are independent queues for each server, this dependence is very strong. If, on the other hand, two servers serve a single queue, a customer's waiting time is only partially dependent on each preceding customer's processing time. If the preceding customer requires a long processing time, chances are high that the other server will become available to serve the waiting customer before the first customer has been served. This situation not only reduces our customer's waiting time but also makes it less dependent on the processing times of preceding customers. This insight leads us to the following important observation: *Pooling available safety capacity to serve all customers improves the performance of a service process*. Note the similarity of the benefits of pooling arrivals to those of consolidating safety stocks that we observed in Chapter 7. There, we saw that centralization (pooling) of inventories improves customer service level for the same total inventory investment. Likewise, we have just seen here that pooling safety capacity improves service in terms of reduced queue length and waiting time for the same investment in capacity. Just as pooling inventories reduces total demand variability and the probability of stockouts, combining queues reduces variability in arrivals and leads to reduced waiting times.

In addition to airport security checking, single lines ("serpentine" or "snake" lines) are observed in passenger arrivals in U.S. immigration and customs, post office, department of motor vehicles, banks, hotel lobbies, etc. Often, arrivals do not have to physically stand in one line; each takes a number and the next available server calls numbers on an FCFS basis, as at the deli counter of a supermarket. Similarly, as we saw in call centers, customers call a single number and are kept on hold to wait for the next available agent. In fast food restaurants, such as McDonalds, there may be separate lines for individual registers, but when an order taker becomes free, he or she can take the next customer in line, effectively merging the lines for service.

Sometimes, combining waiting lines into a single line may be impractical due to space limitations and the physical layout of the facilities, as in the case of supermarket checkouts, where a separate line for each checkout clerk may be necessary. Moreover, pooling arrivals is beneficial only if they all have similar processing requirements, although their actual processing times display stochastic variability. If, on the other hand, arrivals are heterogeneous in terms of their processing requirements, serving them with the same pool of resources increases variability in their processing times which may offset the advantages of reduced variability in arrival times due to pooling. Moreover, serving heterogeneous customers requires the resource pool to be multiskilled and flexible. Thus, having a single line for all bank transactions would require each agent to be knowledgeable in all aspects of banking. It may, instead, be better to separate customers according to their specific needs such as deposit/withdrawal, loan applications, account opening/closing, credit card services, etc.

8.7.2 Segregating Arrivals with Specialized Resources

If arrivals differ systematically in terms of their processing requirements or waiting costs, it may be better to segregate and process them separately, because it will reduce variability in processing times within each class. Also the servers could be trained and specialized to serve different types of customers more efficiently, which would reduce their average processing times as well as its variability, again improving the overall performance.

For example, in an experiment at the security check point of Chicago's Midway airport, passengers were directed to one of the three lines based on their experience: beginners (including families), intermediate, and advanced travelers, depending on their experience level. The result was to expedite overall processing and reduce the

average wait time of 10 minutes down to 5 minutes. Similarly, at call centers in banks, airlines, hospitals, or businesses, each caller is directed to separate extensions, depending on their specific requirements. Given different arrival rates and costs of waiting, separate check-in counters for business class and economy class passengers makes sense, while each class has a single line feeding into a pool of agents. In supermarkets, separate checkout counters for customers with 10 items or less results in faster checkout for such customers and reduced overall variability in processing times for all. Thus, the basic rule is that heterogeneous classes of customers should be segregated for processing, while homogeneous classes of customers should be pooled together.

Another approach to reducing the overall cost of waiting is to process heterogeneous customers according to some priority rather than the first-come-first-served rule. In a hospital emergency room, patient arrivals are checked in by a triage nurse on the FCFS basis, who then determines the acuity of the patient's condition (emergent, urgent, and nonurgent)—corresponding to the patient's cost of waiting—and the order of treatment. Naturally, a patient with a heart attack would get priority over one with a broken leg. In some cases, customers can "jump" the queue by paying a price. For example, while priority boarding and deplaning of business class passengers is common, at the U. S. passport office, patent office, amusement parks, and even in a few temples of worship in India, one could pay extra fee to switch to a separate, shorter line by paying a premium price. At amusement parks such as Walt Disney World, Fast Pass allows visitors to skip lines at popular rides. Clearly, the priority service discipline reduces the priority customers' flow time and improves the service provider's revenues by extracting more consumer surplus. Moreover, if the fraction of priority customers is small, the impact on waiting by the rest of the customers will be minimal. For example, letting disabled passengers or those with small children board the plane first helps them without increasing the waiting time for the rest significantly. With limited resources in emergency rooms and intensive care units (ICUs), some hospitals in fact turn away patients who do not need such intense care in order to be able to treat those who really need it. Thus, priority discipline and admission control may in fact improve the overall process performance. However, it may also raise moral and ethical issues: Should rich people have an advantage of shorter waits? Can a hospital refuse anyone emergency care?

In general, the key to improving process performance is to collect sufficient information about customer arrivals in terms of their processing requirements and their tolerance for waiting, so that appropriate type and level of processing capacity, and also the priority discipline can be tailored to optimize the overall process performance. The challenge is to determine correct classification of arrivals into optimal number of segments, each processed by specialized servers. The goal is to balance advantages of aggregation and segregation, with appropriate level of resource flexibility and specialization necessary to provide fast, consistent service. The triage system in a hospital emergency room is an example of such a hybrid process arrangement. Patients with widely differing problems are segmented by a triage nurse into distinct, more or less homogeneous, classes treated by specialized doctors. Similarly, upon arrival in a bank, the customer is directed to wait for a loan officer, investment consultant, or account manager, depending on his or her need.

8.8 PERFORMANCE IMPROVEMENT LEVERS

In the preceding sections, we saw how process performance (in terms of customer delays and queues, and low throughput due to blocking and abandonment) suffers because of high arrival rate, insufficient buffer and processing capacity, and unsynchro-

nized variability in arrival and processing times. Therefore, the key levers to improve process performance along these dimensions are the following:

1. Decrease capacity utilization (or increase safety capacity) either by
   (a) decreasing the arrival rate or increasing the unit processing rate, or by
   (b) increasing the number of servers
2. Decrease variability in customer interarrival and processing times
3. Synchronize the available processing capacity with demand
4. Increase buffer capacity to reduce blocking arrivals
5. Pool capacity across homogeneous arrivals and separate heterogeneous arrivals

In the following section, we outline concrete managerial actions for implementing these five levers.

8.8.1 Capacity Utilization Levers

The queue length formula shows that decreasing capacity utilization will decrease delays and queues. To decrease capacity utilization $u = R_i / R_p$ (or increase safety capacity $R_s = R_p - R_i$), we can either decrease the inflow rate R_i or increase the processing rate R_p. We discuss managerial actions for achieving each.

Manage Arrivals In manufacturing operations, reducing the arrival rate R_i requires scheduling procurement of raw materials only as needed and hence in small quantities, which minimizes the input inventory. In Chapter 6, we have already seen the benefits of batch size reduction on average inventories and flow times. In Chapter 10, we will see other implications of this just-in-time procurement strategy.

In customer service operations, we have only limited control over customer arrivals. Moreover, if the mission of a service process is to attract and serve customers or if customers pay for service, it does not make sense to reduce their arrival rate. However, it may be possible to shift customer arrivals across time to make them less variable with lower peak rates through reservation systems for scheduling arrivals, differential pricing as incentive to reduce peak demand, offering alternative types of service, or even imploring them to shift demand over time. We have already seen in the airport security example that staggering the flight schedule led to reduced arrival rate at the X-ray scanner. Differential pricing also provides incentives to customers to shift their demand away from peak periods. Reduced prices for matinee shows, happy hour and early-bird dinner specials in restaurants, off-season rates for hotels and resorts, and lower telephone and electric utility rates during off-business hours are some examples. Similarly, by encouraging customers to use online services such as checking balances, automatic monthly deposits of checks and withdrawals of utility bills, and placing ATMs in dispersed convenient locations, banks aim to reduce the need for customers to come to the bank, thereby reducing the arrival rate and congestion. Permitting renewal of driver licenses by telephone and car registration by mail reduces the need to visit the Department of Motor Vehicles (DMV), reducing the arrival rate at the facility. Growth of online shopping has provided flexibility for customers to order and retailers to supply, leading to reduced traffic and congestion in their brick and mortar stores (although adding to shipping time and cost). Finally, organizations may also try appealing to customers to call during off-peak periods, do their holiday shopping early, file income tax returns early and online, and so forth. In energy consumption, "smart meters" show electricity usage patterns and inform consumers a cheaper time to do laundry, which can reduce the peak demand. *Since these actions try to influence the demand pattern, they are called* **demand management** *strategies*. The overall objective is to try to reduce the arrival rate during peak periods of congestion as well as to reduce variability in inflows, resulting in reduced delays and queues.

In fact, if the queues are visible to arrivals, demand management may be self-enforcing, as arrivals seeing long queues may decide not to join. Modeling this phenomenon would require the arrival rate to be dependent on the queue length, and is beyond the scope of this book. It also reminds us of "Yogi" Berra's famous quote about a restaurant being so crowded that nobody goes there anymore!

Managing Capacity Capacity utilization $u = R_i / R_p$ can also be reduced (and safety capacity increased) by increasing the average processing rate R_p. As we saw in Chapter 5, the processing capacity $R_p = c/T_p$ can be increased by either increasing the number of servers c or decreasing the average processing time T_p, and both alternatives involve costs. Processing time reduction can be achieved through process simplification, worker specialization, and the use of high technology (e.g., using bar codes and electronic scanners at checkout counters, or providing computerized patient records in a medical clinic). In McDonald's drive-through, clean, clear menu board and speaker clarity are emphasized to improve speed as well as accuracy of order processing. Parallel processing of nonoverlapping activities or customers leads to compressing their processing times. Starbuck's baristas often make multiple coffee drinks simultaneously to save time, although management also worries about the customer perception of the quality of the beverage served. Finally, designing the service to include some preprocessing (as preregistration at hospitals or printing boarding passes on line) and customer participation (as with a self-service salad bar) are aimed at reducing the server processing requirements and time.

Generally, the cost of providing sufficient capacity to completely eliminate delays cannot be justified on economic grounds, nor may it be possible due to processing variability. The role of the operations analyst, therefore, is to design a service process that achieves an acceptable balance between the operating costs and the delays suffered by customers. Economics of optimal capacity investment decisions should take into account the costs and benefits of adding capacity, as we saw in Section 8.4.

8.8.2 Variability Reduction Levers

As discussed before, variability in interarrival and processing times can be expressed in terms of the coefficients of variation (standard deviation as a fraction of the mean) of their probability distributions. From the queue length formula in Section 8.3, note that the average queue length (and hence waiting time in queue) is directly proportional to the sum of the squares of the two coefficients of variation of interarrival and processing times.

Hence, one lever to decrease the average queue length and waiting time is to reduce variability in arrival and processing times. By planning for more regular arrival patterns, one can reduce the variability in arrival times. In manufacturing, for example, it means choosing more reliable suppliers who will deliver more consistently within narrower time windows. (Recall from Chapter 7 that less variability in the delivery lead time also results in reduced safety inventory.)

Even in service operations where there is only limited control over customer arrivals, it is possible to make arrivals more predictable through scheduling, reservations, and appointments. For example, airlines, hotels, medical offices, and restaurants try to match available capacity with uncertain demand through reservations and appointments. However, because of late arrivals and no shows, variability in arrival times cannot be eliminated completely.

To reduce variability in processing times, we must first understand its source. In some instances, a common resource is used for producing a variety of products, each requiring very different processing times. In such situations, we can reduce the processing time variability by limiting product variety or specializing resources to perform only a narrow range of processing. Examples include standard express meal packs in

fast food restaurants, specialized teller windows at banks, and separate extensions for different types of telephone calls. In some cases, processing times are variable because of a lack of process standardization or insufficient workforce training in standard operating procedures. The solution then is to standardize the process to reduce worker discretion and to set up worker training programs in these standardized procedures. Toyota, for example, defines an exact sequence of activities for each workstation, resulting in a reduction in the variability in processing times as well as in the average processing time. Similarly, because of learning by doing, more experienced workers tend not only to process faster but also do so with higher consistency in terms of the processing time (as well as output quality). Therefore, any managerial actions and incentives aimed at maintaining a stable workforce and low turnover rate will lead to shorter and more consistent processing times and better process performance in terms of shorter queues and delays.

In spite of these strategies, it is impossible to eliminate all sources of variability. Banks, for example, cannot force customers to come in at regular intervals; after all, each customer decides when to go to the bank independently of others. Likewise, banks cannot eliminate processing time variability completely because different customers have different transaction needs, and all these cannot be standardized. In fact, the primary virtue of make-to-order processes is their ability to provide customization. Thus, given the presence of unavoidable variability in inflow and processing times, managers must deal with it by investing in some safety capacity albeit at a higher cost.

8.8.3 Capacity Synchronization Levers

As we saw in Section 8.2, queues build up not just because of variability in inflows and processing times but also because these two sources of variability are unsynchronized. In Section 8.3, the queue length formula as well as the exponential model assumed that arrival times and processing times are independent random variables. Therefore, as we saw in the preceding two subsections, reducing delays and queues requires reducing these two types of variability or reducing the capacity utilization by adding safety capacity. In this section, we consider strategies for synchronizing capacity with variable demand to reduce delays and queues.

In general, synchronization of supply and demand requires managing either the supply (capacity) or the demand (arrival rates). Now, continually adjusting capacity in response to demand may not be economical or even feasible. For example, we cannot change the number of rooms in a hotel or tables in a restaurant as guests come in. In such cases, we must manage demand. As we have seen above, off-season hotel rates and differential pricing are economic incentives designed to even out demand so as to align it with the available supply.

In the short term, personnel adjustments are easier to implement than adjusting capital resources. Strategies to alter capacity with demand are observed at checkout counters in supermarkets, fast food restaurants, and drugstores. There, the store managers open and close checkout counters depending on the number of customers waiting in line. Personnel involved in less time-critical tasks (such as replenishing shelves or cleaning up aisles) are often shifted to staff the newly opened counters, thereby temporarily increasing the processing capacity in busy periods. Once the queue diminishes, these temporary servers return to their original duties. This capacity adjustment strategy is also common in call center operations, where backroom employees or even management personnel answer telephone lines when queues get too long. At Walt Disney World, video cameras and digital maps are used to track waiting lines at popular rides with the goal of dispatching help as gridlocks form. As we will study in Chapter 9, a "control limit policy" specifies when to adjust the capacity (up or down) dynamically

over time depending on the queue length. This capacity adjustment strategy also illustrates the advantages of pooling the available total capacity across different tasks, but it requires that all personnel are trained to perform different tasks. It illustrates the importance of resource flexibility in reducing the permanent capacity requirements, as we will see in Chapter 10. Finally, servers often tend to work faster as the queues get longer in an attempt to synchronize the processing rate with the arrival rate. In any case, these short-term capacity adjustments enable synchronization of capacity with demand and improve the process performance by reducing customer waiting times.

In a somewhat longer time frame, synchronization of capacity with demand is easier to implement. In several businesses—for example, call centers, banks, and fast food restaurants—demand for service varies by the time of the day and the day of the week. Much of this seasonal variability can be anticipated with a high degree of accuracy. The process managers can then plan the required capacity (personnel) to match the forecasted demand. McDonald's, for instance, plans the required number of personnel in 15-minute intervals by scheduling their shifts and breaks and by using part-time workers.

Finally, in manufacturing operations, if the output of one workstation is an input into the next, managers can synchronize the arrival and processing rates by limiting the size of the buffer that is allowed to build up between the two workstations. The feeder workstation is then forced to stop once its output buffer is full. The synchronization decreases the in-process inventory and waiting time but also results in some loss of throughput, as we will see in Chapter 10.

8.8.4 Buffer Capacity Levers

As we have seen in call center operations, adding more telephone lines increases the buffer capacity, allows more calls to come through, and increases the process throughput, albeit at an additional cost. It may also increase the waiting time of customers who do get in. These conflicting factors and their economic impact need to be assessed in deciding the optimal buffer capacity.

Input buffer in a restaurant may correspond to a cocktail lounge where customers can order drinks and appetizers, study the menu, and perhaps even order the meal while waiting for the table. As we will see, this strategy not only mitigates the customer's displeasure with waiting but in fact generate an additional source of revenue from the waiting customers.

8.8.5 Pooling and Segregation Levers

As we saw in Section 8.7, merging of queues of similar customers leads to improved utilization of the available capacity. Pooling is often implemented in banks, post offices, departments of motor vehicles, where a single line feeds into the server pool, instead of a separate line for each server. Similarly, in call centers, all customers call one number and are then routed to the next available agent. Note that if arrivals can see and switch between queues, then having separate queues for different servers is equivalent to a single queue that is served by the entire pool. Thus, for example, in restaurants such as McDonald's, each cash register has a separate queue of customers waiting to place orders. However, if a server becomes idle, he/she takes orders from customers in adjacent lines (although perhaps not necessarily in the FCFS fashion).

Note, however, that pooling different types of customers increases variability in processing times and requires that the servers have sufficient flexibility to be able to process a variety of jobs. The cost of cross-training required must be weighed against the benefits of pooling. Similarly, we have seen above that specialization reduces the average processing time as well as its variability and may improve the service quality. Therefore, segregating customers according to their processing requirements will result

in reduced waiting times and queues within each class. Supermarkets keep special checkout counters for customers with fewer items, which reduces the mean as well as variability in their processing times, thus reducing the average overall wait.

The key to allocating the available capacity is to collect information about processing requirements of customer arrivals, so that customers can be classified into homogeneous classes and each class can assigned to a separate pools of servers. The result will be reduced average and variability in processing times within each class due to server specialization.

8.9 MANAGING CUSTOMER PERCEPTIONS AND EXPECTATIONS

In most service and other make-to-order operations, waiting is a fact of life with significant economic and behavioral implications. The act of waiting has disproportionately large impact on the customers' perception of the total service experience. Long waits sour customers' assessment of an otherwise excellent service. Unfortunately, in practice, waits, queues, and congestion cannot be eliminated completely. However, their adverse impact on customer satisfaction can often be mitigated through behavioral strategies. Various approaches are detailed in Maister (1985), each dealing with the management of customer perception, expectation, and experience. The goal is to make the customer less sensitive to waits, thereby reducing the cost of waiting in terms of customer dissatisfaction and loss of future business.

Provide Comfort Comfortable seating, well decorated, well lit surroundings, staffed by cheerful friendly, helpful servers makes customers wait less unpleasant. Background music while waiting on hold is another example of an attempt to reduce the cost of waiting (although wrong type or loud music may in fact have a negative effect).

Provide Distraction Occupied wait seems less unpleasant than idle wait. Occupying customers with some activity distracts their attention away from the unpleasant act of waiting. Hotel guests have been found to complain less about waiting for elevators near which mirrors are installed; self admiration seems to be the best form of diversion! TV monitors in waiting areas of car repair shops and hospital emergency rooms, and providing video games, entertainment, and amusement to customers waiting for rides in theme parks such as Disney World are some examples for filling waiting time. Restaurants often keep customers occupied by letting them look at menus, play puzzles, order hors d'oeuvres and drinks while waiting for tables (which has the added advantage of generating extra revenues and reducing the serving times required once guests arrive at tables). Some restaurants provide customers with pagers and inform them when a table becomes available, allowing them to wander around and occupy themselves with other activities while waiting. Finally, Disney amusement parks are well-known for entertaining customers while waiting for rides.

Provide Information Several studies have shown that uncertainty about the length of wait ("blind waits") makes customers more anxious and impatient. They tolerate waits better if they are informed of the expected waiting times when they join the queue, with frequent updates. Managers of amusement parks found that number of customer complaints dropped significantly after the park authorities started displaying the expected waiting times for popular rides.

Provide Explanation Explaining why the wait is unavoidable and what is being done to reduce it lessens its impact. When customers sense that the management is aware of the customers waiting and is doing something about it, it creates the empathy factor and improves their relationship with the service provider. Thus, airline passengers are more willing to accept delays in takeoffs and landings as unavoidable if they

are weather related, or essential for passenger safety in case of equipment malfunction. The goal is to align the incentives of the customer and the service provider.

Manage Expectations Customers are often willing to wait longer if the service itself is time consuming. At supermarkets, customers with full carts are willing to wait longer than those purchasing only a few items. Sometimes, pessimistic estimates are provided so customers are pleasantly surprised when the actual wait turns out to be less than the announced period (although, if the announced period is too long, the customer may choose to abandon the queue!). By setting the customers' expectations of delay low and then exceeding those leaves a positive memory of the experience. Airlines often pad their schedules, so when a flight arrives "ahead of schedule" passengers are pleasantly surprised.

Ensure Fairness Customers often complain more readily if they perceive that later arrivals have been served first (even if their own wait is not long). Conversely, if customers are served in the order of their arrival, they accept it as a fair practice and usually elicit fewer complaints even if their waits are long. One of the virtues of pooling arrivals in a single waiting line is its perceived sense of fairness; with separate waiting lines, the other line always seems to move faster than ours! Although providing priority service to customers paying a premium price may be economically rational, it may appear unfair to the rest who perceive that the rich get preferential treatment. To disguise the appearance of inequity, priority customers should be processed discreetly.

Finally, although waiting is generally considered detrimental to customer satisfaction, sometimes quick service and short waits may be perceived to imply lower quality experience! For instance, in contrast to a fast food restaurant, fine dining experience would involve a long, leisurely meal, prepared and served in a relaxed atmosphere. The same customer may have different needs, depending upon whether she wants to have a quick bite to eat for lunch or celebrate a special occasion with a friend. Accordingly, operating strategies should also be tailored appropriately to emphasize response time and quality.

Thus, managing customer expectations and perceptions of the wait could be just as important a lever as reducing the actual waiting time itself.

Summary

In this and in the previous chapter, we considered the fundamental problem of matching the available supply of products and services with variable demand in managing process flows. In Chapter 7, we saw how safety inventory can be used to achieve this objective in make-to-stock processes, while in this chapter we concentrated on the role of safety capacity in make-to-order processes. In these processes, we have seen that flow units may have to wait in input buffers if resources required for processing them are not immediately available, which increases flow times and input buffer inventories.

In particular, in this chapter we focused on managing service operations where customer arrivals have to wait for servers, resulting in delays and queues that lead to customer dissatisfaction. We saw that waiting occurs because of (1) high capacity utilization, which may be due to high inflow rate or low processing capacity, (2) high variability in interarrival and processing times, and (3) lack of synchronization between available capacity and variable demand. The queue length formula gives an approximation to the average number of customers waiting in a queue as a function of utilization, the number of servers, and the variability parameters. Safety capacity is a measure of excess capacity in the process available to handle customer inflows. Appropriate managerial levers for reducing the cost of waiting and long lines, therefore, include (1) managing safety capacity (by increasing the process scale through more servers and/or speed through process improvement and by pooling safety capacity), (2)decreasing variability in inflows (by using reliable suppliers, better forecasts, reservations, and appoint-

ments) and in processing (by employing standardized operating procedures, better training, and specialized servers), (3) improving synchronization of capacity with demand (by scheduling, peak-load pricing, and use of part-time workers), (4) increasing buffer capacity to reduce blocking and increase throughput, and (5) separating arrivals by their processing requirements (through information technology), and pooling capacity across similar customers. Capacity investment decisions include investments in buffer capacity and processing capacity. They should consider the trade-offs between the cost of capacity, the cost of waiting, and the potential lost sales due to abandonment and busy signals (as appropriate).

While promising flow times to customers, however, variability in flow time needs to be considered. Depending on the desired service level, which measures the confidence with which the process manager wishes to meet the due date, a safety time should be added to the promised duration. As long as there is underlying variability in the arrival and service processes, the safety time will increase with utilization.

Finally, in addition to reducing the queues and waits, customers' perceptions of actual waiting should also be managed. This can be achieved in several ways, including making their waits more comfortable, distracting their attention on the act of waiting by entertaining them, explaining the reasons for their wait, and somewhat overstating the wait involved so that customers are pleasantly surprised when the actual wait turns out to be less than announced.

Key Equations and Symbols

(Equation 8.1) $R_p = \frac{c}{T_p}$

(Equation 8.2) $R = \min(R_i, R_p)$ (throughput without blocking and abandonment)

(Equation 8.3) $u = \frac{R}{R_p}$

(Equation 8.4) $R_s = R_p - R$

(Equation 8.5) $T = T_i + T_p$

(Equation 8.6) $I_i = R \times T_i$

(Equation 8.7) $I_p = R \times T_p$

(Equation 8.8) $I = R \times T$

(Equation 8.9) $u = \frac{I_p}{c}$

(Equation 8.10) $I_i = \frac{u^{\sqrt{2(c+1)}}}{1-u} \times \frac{C_i^2 + C_p^2}{2}$

(Equation 8.11) $T = \frac{1}{R_p - R_i} = \frac{1}{R_s}$

(Equation 8.12) $R = \min[R_i(1 - P_b)(1 - P_a), R_p]$

(Equation 8.13) $\text{Prob}(\text{Total time in Process} \leq T_d) = SL$

(Equation 8.14) $T_s = T_d - T$

where

R_p = Processing rate
c = Number of servers
T_p = Processing time
R = Throughput rate
R_i = Inflow (arrival) rate
u = Capacity utilization
R_s = Safety capacity
T = Average total time in process
T_i = Average time in input buffer
I_i = Average number of flow units waiting in input buffer
I_p = Average number of flow units in process
I = Average total number of flow units in the system
C_i = Coefficient of variation in interarrival times
C_p = Coefficient of variation in processing times
P_b = Probability of blocking
P_a = Probability of abandonment
T_s = Safety time
T_d = Promised duration
SL = Service level

Key Terms

- Abandonment
- Blocking
- Buffer capacity
- Coefficient of variation
- Demand management
- Promised duration
- Inflow rate
- Interarrival time
- Pooling capacity
- Processing rate
- Processing time
- Proportion abandoning
- Proportion blocked
- Queue length formula
- Safety capacity
- Safety time
- Service order discipline
- Service rate
- Single-phase service process
- Stability condition
- Stochastic variability
- Throughput delay curve

Discussion Questions

8.1 A fundamental problem in operations management is that of matching supply and demand. What possible strategies can process managers use to address this problem?

8.2 Why are different strategies needed to manage make-to-stock and make-to-order processes?

8.3 In service operations such as supermarkets and medical clinics, process managers strive to make sure that they have sufficient resources on hand to process arriving customers. In spite of this effort, why do we often experience long lines?

8.4 What is the effect of limited buffer capacity on the number of customers who cannot get in and the waiting time of those who do get in?

8.5 In organizing resources to meet variable demand, service process managers can either pool resources so that each resource unit is available for processing any customer, or they can assign specific resources to specific types of customers. Discuss the pros and cons of each strategy and state which strategy you would recommend under what circumstances.

8.6 Discuss and contrast the following three statements:
- "The goal of every process manager should be to satisfy as many customers as possible as quickly as possible."
- "The goal of every process manager should be to minimize queues and inventories."
- "The goal of every process manager should be to maximize product availability."

8.7 In this chapter, we emphasized strategies for improving the process performance in terms of the average flow time and average inventory. Give examples in which it may be inadequate to consider only the average values of these measures.

8.8 In this chapter, we considered mostly *quantitative* measures of process performance in operational terms (such as flow time and inventory) as well as economic terms (including operating revenues and costs). Give five examples of strategies that would improve the perception of the process performance in *qualitative* terms by reducing the psychological impact on customer satisfaction.

Exercises

8.1 A call center has a total of 12 telephone lines coming into its customer service department, which is staffed by 5 customer service representatives. On average, 2 potential customers call every minute. Each customer service representative requires, on average, 2 minutes to serve a caller. After great deliberation, management has decided to add another line, increasing the total to 13. As a result, the call center can expect the following:

a. The proportion of potential customers getting a busy signal will
increase
decrease
be unchanged

b. Average flow time experienced by customers will
increase
decrease
be unchanged

c. Average utilization of customer service representatives will
increase
decrease
be unchanged

8.2 A mail-order company has one department for taking customer orders and another for handling complaints. Currently, each department has a separate telephone number. Each has 7 telephone lines served by 2 customer service representatives. Calls come into each department at an average rate of 1 per minute. Each representative takes, on average, 1.5 minutes to serve a customer. Management has proposed merging the two departments and cross training all workers. The projected new department would have 14 telephone lines served by 4 customer service representatives. As process manager, you expect the following:

a. The proportion of callers getting a busy signal will
increase
decrease
be unchanged

b. Average flow time experienced by customers will
increase
decrease
be unchanged

8.3 Entrepreneur John Doe has just founded Pizza-Ready, which will accept pizza orders for pickup over the phone. Pizza-Ready's strategy is to compete with established pizza restaurants by offering superior, fresh, made-to-order deep-dish pizza and excellent service. As part of his advertising campaign, Doe will publish an ad stating, "If your pizza is not ready in 20 minutes, that pizza plus your next order are on us." Doe has done extensive research on the pizza making process and knows that all fresh deep-dish pizzas require 15 minutes of oven time and 2 minutes of preparation. Moreover, as part of its excellent service, Pizza-Ready will accept orders whenever customers place them, and a marketing study estimates that Pizza-Ready can count on an average demand of 20

pizzas per hour. Doe, therefore, has ordered five pizza ovens, each of which is able to bake one pizza at a time. Doe is now looking for a silent partner to help carry the financial burden of his start-up company. Given the structure of this business, a potential partner has asked you whether Pizza-Ready would be a profitable investment. What would you recommend, and why?

8.4 M.M. Sprout, a catalog mail order retailer, has one customer service representative (CSR) to take orders at an 800 telephone number. If the CSR is busy, the next caller is put on hold. For simplicity, assume that any number of incoming calls can be put on hold and that nobody hangs up in frustration over a long wait. Suppose that, on average, one call comes in every 4 minutes and that it takes the CSR an average of 3 minutes to take an order. Both interarrival and activity times are exponentially distributed. The CSR is paid $20 per hour, and the telephone company charges $5 per hour for the 800 line. The company estimates that each minute a customer is kept on hold costs it $2 in customer dissatisfaction and loss of future business.

a. Estimate the following:
- The proportion of time that the CSR will be busy
- The average time that a customer will be on hold
- The average number of customers on line
- The total hourly cost of service and waiting

b. More realistically, suppose that M.M. Sprout has four telephone lines. At most, therefore, three callers can be kept on hold. Assume, too, that any caller who gets a busy signal because all four lines are occupied simply hangs up and calls a competitor. M.M. Sprout's average loss, in terms of current and potential future business, is $100 per frustrated caller. Estimate the total cost of the following:
- Providing service
- Waiting
- Average hourly loss incurred because customers cannot get through

c. Suppose that M.M. Sprout is considering adding another line in order to reduce the amount of lost business. If the installation cost is negligible, can the addition of one line be justified on economic grounds? How would it affect customer waiting time?

d. In addition to adding another line, suppose M.M. Sprout wants to hire one more CSR to reduce waiting time. Should the firm hire another CSR?

***8.5** Heavenly Mercy Hospital wants to improve the efficiency of its radiology department and its responsiveness to doctors' needs. Administrators have observed that, every hour, doctors submit an average of 18 X-ray films for examination by staff radiologists. Each radiologist is equipped with a conventional piece of viewing equipment that reads one film at a time. Because of complications that vary from case to case, the actual time needed for report preparation is exponentially distributed with a mean of 30 minutes. Together, the cost of leasing one piece of viewing equipment and each radiologist's salary is $100 per hour. Although it is difficult to put a dollar value on a doctor's waiting time, each doctor would like to get a radiologist's report within an average of 40 minutes from the time the film is submitted.

a. Determine the number of radiologists that the hospital should staff in order to meet doctors' requirements regarding job flow time. Compute the resulting hourly cost of operating the radiology department.

b. The hospital could also change its diagnostic procedure by leasing more sophisticated X-ray viewing devices. Administrators estimate that the new procedure would reduce a radiologist's average film-processing time to 20 minutes. At the same time, however, higher equipment rental and salaries for additional support personnel would boost the hourly cost per radiologist to $150.

Determine the number of radiologists that the hospital should staff under this new arrangement. Would the new arrangement be economically advantageous?

8.6 First Local Bank would like to improve customer service at its drive-in facility by reducing waiting and transaction times. On the basis of a pilot study, the bank's process manager estimates the average rate of customer arrivals at 30 per hour. All arriving cars line up in a single file and are served at one of four windows on a first-come/first-served basis. Each teller currently requires an average of 6 minutes to complete a transaction. The bank is considering the possibility of leasing high-speed information-retrieval and communication equipment that would cost $30 per hour. The new equipment would, however, serve the entire facility and reduce each teller's transaction-processing time to an average of 4 minutes per customer. Assume that interarrival and activity times are exponentially distributed.

a. If our manager estimates the cost of a customer's waiting time in queue (in terms of future business lost to the competition) to be $20 per customer per hour, can she justify leasing the new equipment on an economic basis?

b. Although the waiting-cost figure of $20 per customer per hour appears questionable, a casual study of the competition indicates that a customer should be in and out of a drive-in facility within an average of 8 minutes (including waiting). If First Local wants to meet this standard, should it lease the new high-speed equipment?

***8.7** Since deregulation of the airline industry, increased traffic and fierce competition have forced Global Airlines to reexamine the efficiency and economy of its operations. As part of a campaign to improve customer service in a cost-effective manner, Global has

focused on passenger check-in operations at its hub terminal. For best utilization of its check-in facilities, Global operates a common check-in system: passengers for all *Global* flights queue up in a single "snake line," and each can be served at any one of several counters as clerks become available. Arrival rate is estimated at an average of 52 passengers per hour. During the check-in process, an agent confirms the reservation, assigns a seat, issues a boarding pass, and weighs, labels, and dispatches baggage. The entire process takes an average of 3 minutes. Agents are paid $20 per hour, and Global's customer relations department estimates that for every minute that a customer spends waiting in line, Global loses $1 in missed flights, customer dissatisfaction, and future business.

a. How many agents should Global airlines staff at its hub terminal?

b. Global has surveyed both its customers and its competition and discovered that 3 minutes is an acceptable average waiting time. If Global wants to meet this industry norm, how many agents should it hire?

8.8 When customers of Henniker Bank believe a mistake has been made on their account statements, their claims are forwarded to the bank's research department, whose trained clerks carefully research and document the transactions in question. On completing her investigation, a clerk phones the customer with her findings. The research department has three clerks. Each handles claims from a separate geographic district and never works on claims from outside her own district. The average number of complaints arising from each district is the same, 3.5 per week. The clerks are equally experienced and completely process the average claim in 1.2 days. Assume a five-day week.

a. Across all districts, how many claims are waiting to be processed on average? What fraction of claims is completed in less than 10 business days?

b. The bank is considering introducing a new information system that would reduce the standard deviation of the service distribution by 50%, although the mean would remain unchanged. How would your answers to part a change?

***8.9** Burrito King, a new fast-food franchise, has had problems with its drive-through window operations. Customers arrive at an average rate of one every 30 seconds. Current service time has averaged 25 seconds with a standard deviation of 20 seconds. A suggested process change, when tested, results in an average service time of 25 seconds with a standard deviation of 10 seconds. Assume that no customers are blocked or abandon the system.

a. As a result of implementing this change, will the average waiting time in queue increase, decrease, or remain unchanged?

b. As a result of implementing this change, will the average server utilization increase, decrease, or remain the same?

8.10 V.V. Ranger is a seller of industrial products. All purchases by customers are made through call centers where Ranger representatives take orders. Currently, Ranger has over 350 warehouses in the United States, each with its own call center. Customers call one of the call centers and wait on hold until a representative at that call center becomes available. Ranger is evaluating a switching system where customers will call one 800 number from where they will be routed to the first available representative in any of the call centers. If Ranger installs the switching system, will the average waiting time of customers increase, decrease, or remain the same? Explain.

***8.11** Master Karr is a supplier of industrial parts. All orders are received at a call center. The call center has 15 phone lines, so that a maximum of 15 callers may be in the system at a time. Calls arrive at an average of 4 calls per minute. The call center currently has 5 customer service representatives (CSRs). Each CSR averages 1 minute a customer. Master Karr estimates that waiting costs incurred are $1 per customer per minute in terms of phone charges and loss of future business. Also assume that callers who get a busy signal take their business elsewhere, resulting in a loss to Master Karr of $50 per lost call. Assume that callers do not abandon once they enter the system. CSRs are paid $15 per hour.

a. What is the hourly cost to Master Karr of the current configuration of the call center?

b. What is the hourly cost to Master Karr if they decide to hire another CSR? Do you recommend this move?

8.12 BizTravel.com is a travel Web site that recently announced "BizTravel Guarantee," putting money behind customer-service guarantees.

a. One of the items in the BizTravel guarantee states, "If your customer service e-mail is not responded to within two hours, we'll pay you $10." Customers currently send e-mails to service@biztravel.com. The e-mail server of BizTravel equally distributes these e-mails to the specific address of each of the five CSRs. For example, one-fifth of the e-mails are directed to the mailbox of CSR1@biztravel.com, another one-fifth to CSR2@biztravel.com, and so on. Collaborative Inc. has developed collaborative software for customer relationship management that allows the firm to keep all customer service requests in a central mailbox and dynamically route the e-mails based on agent availability. Do you think the software from Collaborative Inc. will help BizTravel meet its customer guarantee better? Explain.

b. Another service guarantee offered is, "If your phone call is not answered within 90 seconds, we'll pay you $10." Peak arrival rate of calls to BizTravel is during the lunch hour from 12 to 1, and averages one customer every minute. A transaction takes on average 5 minutes to service. The manager decides to schedule 5 agents during this period. Do you expect the BizTravel to have to pay out any money?*

***8.13** Drive-through window operations are becoming an increasing source of competitive advantage for the fast-food restaurant business. McBerger's has performed poorly in this area compared to Mandy's, the leader in drive-through operations. The service from a drive-through window is staged. At the first stage, the customer places an order. At the second stage, the customer makes a payment at the payment window. Finally, at the third stage, the customer picks up the order. The time between consecutive customer arrivals is exponentially distributed with an average of 45 seconds. Currently, McBerger's total service time (across three stages) averaged 55 seconds with a standard deviation of 35 seconds. Several new process changes were made. Assume that no customers abandon the system or are blocked after entry in either system (before or after the change).

a. Competitors have experimented with a separate kitchen to service the drive-through orders. When McBerger's implemented this new "plant-within-a-plant" strategy, average service time remained at 55 seconds but with a standard deviation of 25 seconds. As a result of this change, did the average waiting time in queue increase, decrease, or remain the same?

b. McBerger's began testing the installation of a transponder on a customer's windshield that allowed the restaurant to scan the identification of the car. Using this technology, the customers were billed directly instead of paying at the window. As a result of this technology, do you think the average waiting time increased, decreased, or remained the same?

Selected Bibliography

Andrews, B. H., and H. L. Parsons. "L. L. Bean Chooses a Telephone Agent Scheduling System." *Interfaces* 19, no. 6 (November–December 1989): 1–9.

Cachon, G., and C. Terwiesch. *Matching Supply with Demand.* New York: McGraw-Hill/Irwin, 2006.

Chase, R., F. R. Jacobs, and N. Aquilano. *Production and Operations Management.* 10th ed. New York: McGraw-Hill/Irwin, 2004.

Hopp, W. J., and M. L. Spearman. *Factory Physics*, 3rd ed. Chicago: Irwin, 2008.

Kleinrock, L. *Queueing Systems.* New York: John Wiley & Sons, 1975.

Larson, R. C. "Perspectives on Queues: Social Justice and the Psychology of Queuing." *Operations Research* 35, no. 6 (November–December 1987): 895–905.

Maister, D.. "The Psychology of Waiting Lines." In *The Service Encounter: Managing Employee/Customer Interaction in Service Businesses,* ed. J. A. Czepiel, M. R. Solomon, and C. F. Suprenant. Lexington, MA: Lexington Books, 1985.

McClain, J. O., L. J. Thomas, and J. B. Mazzola.. *Operations Management.* 3rd ed. Upper Saddle River, N.J.: Prentice Hall, 1992.

APPENDIX

The Exponential Model with Finite Buffer Capacity

Suppose:

- Interarrival and processing times are exponentially distributed with the mean arrival rate R_i and the mean processing rate $R_p = c/T_p$ (so that the mean interarrival time $= 1/R_i$ and the mean processing time $= T_p$).
- There are c servers, each processing customers joining a single queue in FCFS order.
- Input buffer capacity is K (so that $K + c$ is the maximum number of customers that can be in the system).

Explicit formulas are available for computing performance characteristics of such a system (for details, see Kleinrock, 1975)—in particular, the probability P_n that there are n customers in the system, i.e, total inventory $I = n$ is

$$P_n = \begin{cases} \dfrac{1}{n!}(T_pR_i)^n P_0 & \text{if } 0 \le n < c \\ \dfrac{1}{c!c^{n-c}}(T_pR_i)^n P_0 & \text{if } c \le n \le c + K, \end{cases}$$

where P_0 is the probability that there is no one in the system, which is given by

$$\frac{1}{P_0} = \begin{cases} \displaystyle\sum_{n=0}^{c-1} \frac{1}{n!}(T_pR_i)^n + \frac{(T_pR_i)^c}{c!}\frac{(1 - (T_pR_i)^{k+1})}{(1 - T_pR_i)} & \text{if } R_i \ne R_p \\ \displaystyle\sum_{n=0}^{c-1} \frac{1}{n!}(T_pR_i)^n + \frac{(T_pR_i)^c}{c!}(K + 1) & \text{if } R_i = R_p \end{cases}$$

Note that the blocking probability P_b is the probability that there are $K + c$ customers in the system, that is, $P_b = P_{K+c}$. The capacity utilization is given by

$$u = \frac{R}{R_p} = \frac{R_i(1 - P_b)}{R_p} = \frac{T_pR_i(1 - P_{K+c})}{c}$$

Defining $r = R_iT_p$, we have

$$I_i = \frac{P_0(cr)^c r}{c!(1 - r)^2}\left[1 - r^{K+1} - (1 - r)(K + 1)r^k\right],$$

$$I = I_i + c - P_0\sum_{n=0}^{c-1}\frac{(c - n)(rc)^n}{n!}; \quad T_i = \frac{I_i}{R}; \; T = \frac{I}{R}.$$

From these, other performance measures can be computed by using their definitions along with Little's law. These formulas may also be used when there is no limit on buffer capacity by setting the buffer capacity K to be a very large number (infinity). As mentioned earlier, these formulas have been programmed on a spreadsheet called **Performance.xls** that can be downloaded from the publisher's web site at www.prenhall.com/anupindi.

Chapter 9

Managing Flow Variability: Process Control and Capability

INTRODUCTION

Overhead Door Corporation's founder, C. G. Johnson, invented the upward-lifting garage door in 1921 and the electric garage door opener in 1926. Since then Overhead Door has been a leading supplier of commercial, industrial, and residential garage doors sold through a nationwide network of more than 450 authorized distributors. Overhead Door employs the latest computer-aided design and manufacturing (CAD/CAM) methods for developing and producing doors to exact customer specifications. The company's brand is known for high quality of their products and professional after-sales service and they have built a solid reputation as a premier door supplier, commanding 15 percent share of the market.

At their recent holiday party, Overhead Door employees were celebrating the company's success over the year past. During the many speeches given, executives were congratulating one another for the job well done. However, the sales manager stunned everyone by announcing, "Ladies and gentlemen, I do not wish to spoil your mood, but I have some disturbing news! Lately, I have been talking to some of our major customers, and I have found, much to my surprise, that many of them are less than satisfied with our products and services. In fact, one distributor said to me the other day that our overall quality stinks! Although we think our products are great and that our service is unsurpassed, if

what I am hearing is right, it is only a matter of time before we lose our loyal customer base to the competition such as Clopay, which is working hard to provide newer and better products, cheaper and faster." Typically a messenger of bad news like this would be ignored, ridiculed, or fired, especially given everybody's perception of the company as a successful enterprise. However, the chief executive officer (CEO) of Overhead Door was not an ordinary individual; she was a thoughtful leader with vision, wisdom, and an open mind. So, she asked the sales manager to elaborate further. He explained that several customers that he talked to were unhappy with their door quality in terms of safety, durability, and ease of operation; others were annoyed that their doors cost much more than the competition; and still others complained about the difficulty in getting their orders delivered on time or receiving prompt service when something went wrong with installation or operation. The CEO listened carefully and concluded that it was time to be proactive by identifying and eliminating root causes of customer dissatisfaction with the company's products and services.

The CEO wondered if the sales manager's observations were based on subjective impressions and isolated instances. As a management trainee back in 1990s, she had attended seminars on principles of **total quality management (TQM)** and learned about six sigma quality tools. So in the spirit of "management by fact," she decided that the next logical step should be to collect and analyze some hard data. This would not only provide objective assessment of the customer experience with the company's products and services, but also facilitate taking specific corrective actions based on quantitative scientific evidence rather than on mere intuition, emotion, or hearsay. Accordingly, the CEO formed an interdisciplinary quality improvement team (QIT) comprising the sales manager, production engineer, product designer, material supplier, and service manager. She assigned them the task of collecting and analyzing concrete data on critical performance measures that drive customer satisfaction, with the goal of identifying, correcting, and preventing the sources of future problems. By way of illustration, we will trace the steps that QIT might take to uncover root causes of customer dissatisfaction with the company's products and services.

All products and services display variability in terms of their cost, quality, availability, and response times, which often leads to customer dissatisfaction, as the story above illustrates. In this chapter, we study some graphical and statistical methods for measuring, analyzing, controlling, and reducing variability in product and process performance, with the goal of improving customer satisfaction. In Section 9.1, we discuss how product and process variability affects customer satisfaction. In Section 9.2, we present some simple graphical and statistical methods for measuring, organizing, visualizing, and analyzing this variability. Although the concepts and methods throughout this chapter are applicable to managing variability in any metric of product or process performance—including cost, quality, availability, and response time—for purposes of illustration we will stress quality as the key attribute, since we have already studied the others in detail in the previous chapters.

In Chapters 7 and 8, we analyzed the detrimental effects of variability in supply (lead times and processing times) and demand (order quantities and arrivals) on product and process availability and response time. There we assumed that the statistical nature of variability was stable and known in terms of its probability distribution. Our response to this variability was then a *static* plan of building in safety nets (in the form of safety inventory, safety capacity, or safety time margin) that would absorb some of the variability in the actual supply or demand so as to provide a desired level of performance most of the time.

In practice, however, statistical laws governing variability may themselves be unknown, and, moreover, they may be changing over time. Therefore, in this chapter our emphasis will be on estimating, tracking, and responding to this variability over time. In

particular, in Section 9.3, we study online process control, which involves *dynamically* monitoring the actual performance over time and taking corrective actions in light of observed deviations from the planned performance. Process control involves tracking a key performance measure, comparing it against the expected level of performance, and signaling the need for a corrective action whenever the observed performance deviates excessively from the expected one. In particular, we outline a "control limit policy" that specifies investigation when—and only when—the observed performance measure exceeds certain critical thresholds. We discuss statistical process control (SPC) as a prominent example of such a policy for managing quality, and indicate its application to controlling inventory, capacity, and cost.

The objective of process control is an *internal* one of identifying and eliminating abnormal variability thereby maintaining the process in a stable state of statistical equilibrium that displays only normal variability in its performance. Process capability, in contrast, measures how well the process output meets *external* customer requirements, which is the subject of Section 9.4. It represents accuracy of the process in conforming to customer specifications. In contrast with short-term process control, improving process capability requires long-term investment in resources to reduce normal variability, as discussed in Section 9.5. Finally, in Section 9.6, we show how careful design of products and processes through simplification, standardization, and mistake-proofing minimizes sources of process variability and its impact on product performance. Summary of the managerial levers for designing and controlling product and process variability concludes the chapter.

9.1 PERFORMANCE VARIABILITY

All measures of product and process performance—both external and internal—display variability over time. External measurements (such as customer satisfaction indices, product rankings, and number of customer complaints) vary from one market survey to the next. Quality of instruction, as measured by teacher/course evaluations fluctuates from one term to the next. Business school rankings published by Business Week, Financial Times, and Wall Street Journal vary from year to year. Product and service rankings announced by *Consumer Reports* and J. D. Power and Associates also show variability over time.

In all business processes, flow units vary with respect to their cost, quality, and flow times. No two cars rolling off an assembly line are exactly identical. Even under identical circumstances, the time and cost required to produce and deliver the same product may be quite different. Two different customers (in fact, the same customer on two separate occasions) may assess the quality of a restaurant dining experience quite differently. The cost of operating a department within a company also varies from one quarter to the next. Customers in a bank conducting apparently identical transactions may in fact experience different waiting and processing times. Even with careful ordering, a department store may run out of stock of an item one month and have excess inventory left over the next. Sources of all this variability may be either internal (e.g., imprecise equipment, untrained workers, or lack of standard operating procedures) or external (e.g., inconsistent raw material, supplier delivery delays, changing economic conditions, or changing customer tastes and requirements).

In general, variability refers to a discrepancy between the actual and the expected performance, which usually leads to higher costs, longer flow times, reduced availability, lower quality, and, ultimately, dissatisfied customers. A luxury car that is loaded with options but needs frequent repairs may be judged as inferior to a basic no-frills model that is nevertheless reliable. A highly skilled carpenter who often misses or is late for appointments cannot be recommended for a house remodeling job. A sharpshooter

whose shots are on average centered on the bull's-eye but widely dispersed around it cannot be considered a dependable bodyguard; he is accurate, but imprecise. If, on the other hand, his shots are closely clustered but away from the target, then he is inaccurate but precise, and can readily improve performance by adjusting his aim. As we saw in Chapter 7, a raw material supplier with consistent delivery lead time allows the manufacturer to reduce the safety inventory required to provide a given level of service. Similarly, recall from Chapter 8 that reducing flow time variability lowers the safety time margin required to provide customer service within the promised duration. Recall also that customers often prefer predictable—even if long—waits (e.g., in a restaurant or on the telephone) over uncertain or "blind" waits. Thus, products and processes that display performance variability are generally judged less satisfactory than those with consistent, predictable performance. In short, it is the variability in performance—not just its average—that matters to customers, although businesses tend to pay more attention to the averages. An old adage reminds us that one may in fact drown in a lake that is, on average, only five feet deep!

Although some may enjoy the surprise of the unexpected (as in a surprise birthday party), generally customers perceive any variability in product or service from its expected performance as a loss in value. Japanese quality engineer Genichi Taguchi suggests that this loss be measured by the squared deviation in the actual performance from its target, implying that it increases very rapidly as the actual performance deviates further from the planned one. In fact, product quality or customer satisfaction in general may be defined by the discrepancy between customers' expectation of the product performance and their actual experience with it. It may be due to a gap between the following:

- What the customer wants and what the product is designed for
- What the product design calls for and what the process for making it is capable of producing
- What the process is capable of producing and what it actually produces
- How the produced product is expected to perform and how it actually performs
- How the product actually performs and how the customer perceives its performance

Each of these gaps ultimately leads to customer dissatisfaction or lower quality. In general, we may classify a product as being "defective" if its cost, quality, availability, or flow time varies significantly from their expected values, which leads to dissatisfied customers. From an actionable perspective, it is useful to group these gaps into two classes: (1) gap between customer requirements and product design specifications, and (2) gap between design specifications and actual measurements of the product produced.

Thus, **quality of design** refers to *how well product specifications reflect customer requirements.* From the customer's perspective, product requirements may be defined along several dimensions, such as its features, aesthetics, performance, reliability, durability, serviceability, etc., that customers care about; (see Garvin, 1988). Thus, in making their automobile purchase decisions, customers care about the purchase price as well as styling, safety features, acceleration, gasoline consumption, repair record, and even prestige value. **Quality function deployment (QFD)** is a *conceptual framework for translating customers' functional requirements of a product into its concrete design specifications.* For example, appearance, durability, and ease of operating a garage door must be translated into engineering specifications such as the door material composition, dimensions, and weight. The objective of QFD is to provide a common platform for incorporating the "voice of the customer" into the product design process. Details about QFD may be found, for example, in Hauser and Clausing (1988). House of Quality is routinely used as a first step in many six-sigma projects and is often referred to as Design for Six Sigma (DFSS) in industry.

Quality of conformance, on the other hand, refers to *how closely the actual product conforms to the chosen design specifications.* Thus, a well-made Toyota Camry has a better quality of conformance than a poorly made Lexus, although Lexus has a better quality of design in terms of more power, comfort, and safety features. Quality of design thus refers to what we promise to customers (in terms of what the product or service can do), while quality of conformance measures how well we deliver on the promise (in terms of how it in fact performs). Measures of the quality of conformance, therefore, include such criteria as "number of defects per car" and "fraction of the output that meets specifications." In a bank, for instance, the degree of conformance can be measured by the error rate in check processing and in monthly statements mailed or the percentage of customers who have to wait longer than five minutes for a transaction. In evaluating software services, conformance measurements might include the number of errors per 1,000 lines of code, the percentage of project milestones met on time, the frequency and magnitude of project cost overruns, the number of software program rewrites, or the frequency of system crashes. In an airline, conformance may be measured in terms of the percentage of flights delayed by more than 15 minutes, the number of bags lost per thousand flown, or the number of reservation errors made. The degree of product conformance to design specifications depends on variability in process performance that results in defective products and customer dissatisfaction, eventually leading to loss of reputation, market share, and competitive position. It is therefore critical to measure, analyze, control, and reduce this variability.

9.2 ANALYSIS OF VARIABILITY

In this section, we first present some simple graphical methods for collecting, organizing, and displaying information about variability in product and process performance. Statistics is the science of variability, so we will outline some basic statistical methods for analyzing observed variability. Our goal is to provide diagnostic tools (often called "Six Sigma Quality tools") to help us monitor the actual process performance over time, analyze variability in it, uncover its root causes, eliminate them, and finally prevent them from recurring in the future. Throughout, we will illustrate the key concepts and methods by examining operations at Overhead Door Corporation, whose business motto is "No sale is complete until you are completely satisfied."

Suppose the quality improvement team (QIT) at Overhead Door decides to focus on customers who purchase their standard residential garage doors. They need to know how these customers perceive the total experience of doing business with the company and how it can be improved. Accordingly, they have tried to identify factors that determine customer satisfaction with Overhead Door's products and services and understand how to measure, analyze, and improve them.

9.2.1 Check Sheets

A **check sheet** is simply *a tally of the types and frequency of problems with a product or a service* experienced by customers, as illustrated in Example 9.1.

EXAMPLE 9.1

Suppose the QIT surveyed 1,000 current and past customers, asking them to rate their experiences with each of the following aspects of Overhead Door's products and services:

- Cost of purchasing and maintaining a door
- Response time from ordering a door until its delivery

Type of Complaint	Number of Complaints
Cost	𝍸 𝍸
Response Time	𝍸
Customization	𝍸
Service Quality	𝍸 𝍸 𝍸
Door Quality	𝍸 𝍸 𝍸 𝍸 𝍸

FIGURE 9.1 Check Sheet of Customer Feedback

- Degree of door customization permitted in accommodating individual preferences
- Service quality in terms of order placement experience and after-sales service
- Door quality in terms of its

  Fit and finish
  Ease of operation
  Durability

If customers rate their experience as "unsatisfactory" along any of these dimensions (indicating a gap between their expectation and experience), the pertinent flow unit (customer order) would be considered "defective." The QIT can then compile a check sheet of defectives by type, as shown in Figure 9.1.

9.2.2 Pareto Charts

After counting defects by type, our next step is to determine which defect should be tackled first. All defects are not equal in terms of either their importance or frequency of occurrence. So, given the limited time and resources at our disposal, we should identify and focus only on a few critical ones. We may rank-order types of defects by the frequency of their occurrence or, better yet, according to the frequency weighted by their importance. Problems usually distribute themselves according to the principle of "vital few and trivial many." Thus, the **80-20 Pareto principle** states that *20% of problem types account for 80% of all occurrences.* A **Pareto chart** is simply *a bar chart that plots frequencies of occurrence of problem types in decreasing order.* Example 9.2 illustrates the Pareto chart and its use in the analysis of performance variability.

EXAMPLE 9.2

The record of customer complaints from the check sheet in Figure 9.1 can also be graphed as a column chart in Figure 9.2. As you can see, it identifies door quality as the major problem that the team should address first.

The Pareto chart tells us, for example, that it is better to focus our process improvement efforts first at reducing the tallest bar (door quality) by one-half rather than trying to completely eliminate a short bar (e.g., response time or even cost).

Once the dominant problem is solved, we may collect new data in order to uncover a new tallest bar on which to focus our efforts next. A Pareto chart can thus serve as a dynamic tool for continuous improvement by continually identifying, prioritizing, and fixing problems. Thus, after identifying door quality as the main concern

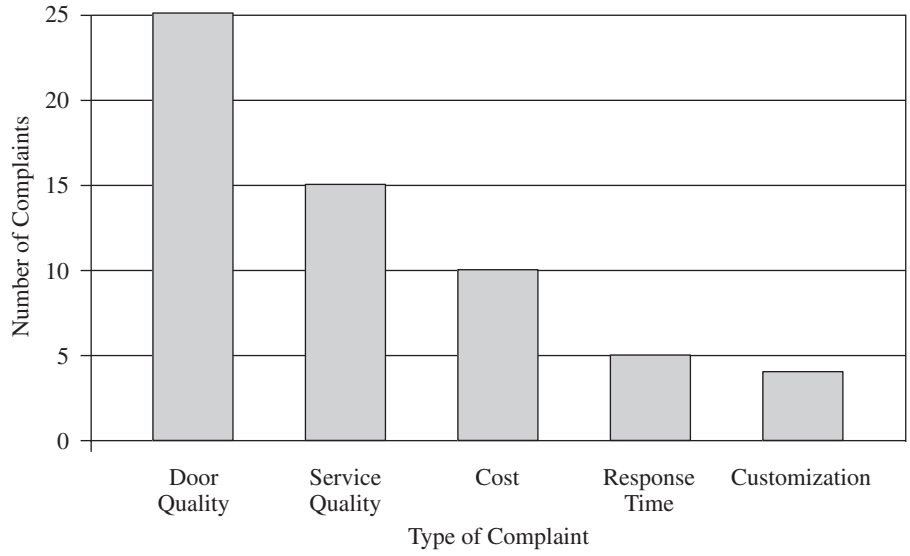


FIGURE 9.2 Pareto Chart of Customer Complaints

voiced by Overhead Door's customers, the QIT could try to pin down exactly what aspects of door quality trouble them most. They could again use a check sheet, this time classifying each defective door according to a new list—poor fit and finish, difficult or unsafe to operate, not durable, and so forth.

Suppose, the second Pareto chart reveals that customers assess door quality first in terms of the ease of operation, followed by its durability. The QIT might then assign an engineering team to determine the factors that contribute to these two main problems. Smooth operation of a garage door depends upon the springs, rollers, tracks, and cables that raise and lower the door. Suppose all this detective work ultimately leads to identifying the weight of a garage door as a critical quality characteristic that affects both problems: If a door is too heavy, it's difficult and unsafe to balance and operate; if it's too light, it tends to buckle and break down frequently or may not close properly. Suppose the design engineers determine that a standard garage door should weigh a minimum of 75 kg. and a maximum of 85 kg., which thus specifies its design quality specification. To determine the quality of conformance, suppose the QIT decides to collect data on the actual weights of 100 standard garage doors sampled randomly from their monthly production of almost 2,000 doors.

9.2.3 Histograms

A **histogram** is *a bar chart that displays the frequency distribution of an observed performance metric.* A preliminary statistical analysis of the performance metric involves summarizing the frequency distribution in terms of its average (which estimates the mean, or expected value at which the distribution is balanced), and the standard deviation, which is a measure of the spread of the distribution around this mean. Example 9.3 illustrates the histogram of door weights.

EXAMPLE 9.3

Suppose five doors from each of the past 20 days' production runs were weighed at two-hour intervals and the weights were recorded as in Table 9.1. As we can see, door

Table 9.1 Garage Door Weight Data

	Day									
Time/Day	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
9 a.m.	81	82	80	74	75	81	83	86	88	82
11 a.m.	73	87	83	81	86	86	82	83	79	84
1 p.m.	85	88	76	91	82	83	76	82	86	89
3 p.m.	90	78	84	75	84	88	77	79	84	84
5 p.m.	80	84	82	83	75	81	78	85	85	80
	Day									
Time/Day	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
9 a.m.	86	86	88	72	84	76	74	85	82	89
11 a.m.	84	83	79	86	85	82	86	85	84	80
1 p.m.	81	78	83	80	81	83	83	82	83	90
3 p.m.	81	80	83	79	88	84	89	77	92	83
5 p.m.	87	83	82	87	81	79	83	77	84	77

weights display variability from door to door within each day's sample as well as between samples from different days. Our ultimate goal is to analyze this variability, determine what action, if any, is necessary to keep it in control, and finally how it can be reduced to improve conformance of actual door weights to design specifications.

The garage door weight data in Table 9.1 can also be displayed more visually as a histogram in Figure 9.3, which shows that 14% of the doors weighed about 83 kg., 8% weighed about 81 kg., and so forth.

We can summarize the entire distribution of door weights in terms of its two key statistics, the overall average weight $\overline{\overline{X}} = 82.5$ kg. and standard deviation $s = 4.2$ kg. (or the variance $s^2 = 17.64$ kg^2.), based on our 100 observations. Thus, $\overline{\overline{X}}$ estimates the average weight of all garage doors produced, and s measures variability in weights from door to door. A higher value of $\overline{\overline{X}}$ indicates a shift in the entire distribution to the right, so that all

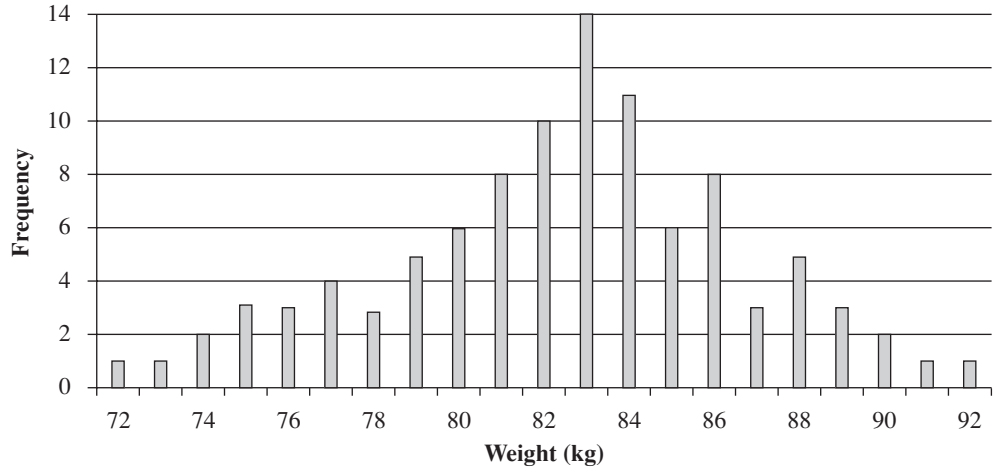


FIGURE 9.3 Histogram of Door Weights

doors produced are consistently heavier. An increase in the value of s means a wider spread of the distribution around the mean, implying that many doors are much heavier or lighter than the overall average weight.

The discrete distribution depicted by isolated bars in Figure 9.3 may be conveniently approximated by a continuous curve, which in this instance would appear as a bell-shaped normal distribution that is symmetric around its mean. Recall the properties of normal distribution that we discussed in the context of safety inventory in Chapter 7 (see also Appendix II). From these properties, we know, for example, that 68.26% of all doors will weigh within ± 1 standard deviation from the average weight—that is, within $82.5 \pm (1)(4.2)$, or between 78.3 and 86.7 kg. Likewise, we know that weights of 95.44% of doors will fall within ± 2 standard deviations from the mean (between 74.1 and 90.9 kg.), and 99.73% of door weights will be within ± 3 standard deviations from the mean (between 69.9 and 95.1 kg.). Standard deviation (or variance) of the output is thus a measure of the variability in the door-making process. A precise, consistent process would produce doors of nearly identical weights, resulting in predictable door quality that is closer to design in terms of its ease of operation and durability.

Similar statistical analysis can be performed on the probability distribution of the response time, cost, and customer experience with the order fulfillment process as well as any other performance metric that may be important to Overhead Door's customers. The key fact is that product and process performance along any dimension varies from one flow unit to another, and we need to measure, analyze, and reduce this variability, with the goal of making performance more predictable and consistent with customers' expectations.

Although a histogram summarizes the overall performance in the aggregate, it does not show how it varies over time, information that is often useful in identifying and reducing overall variability.

Suppose that over the past 20 days there has been a steady upward trend in door weights from an average of 80 to 85 kg., or 0.25 kg. per day. When we aggregate the 20-day data, we may get the same histogram, mean, and variance that we would get if we had made all our observations from the output of Day 10, which also has the average weight of 82.5 kg. What can we say, then, if we had to predict door weights on Day 21 of production? On the basis of the histogram alone, we would think that it will be a random sample from the normal distribution with a mean of 82.5 kg. and a standard deviation of 4.2 kg. However, if we knew the upward trend over time, our estimate of the mean weight on Day 21 should be 85.25 kg.

Thus, if we rely solely on the aggregate performance metric summarized by a histogram, we lose the "time value" of information. In the next section, therefore, we emphasize the importance of tracking process performance over time, which is consistent with our flow perspective throughout this book.

9.2.4 Run Charts

A **run chart** is *a plot of some measure of process performance monitored over time.* It displays variability in the process output across time, which helps us identify structural variability such as trend and seasonality (to be distinguished from stochastic variability due to random noise). A run chart of door weights is illustrated in Example 9.4.

EXAMPLE 9.4

To track variability in door weights over time, we may plot weights of doors sampled at two-hour intervals from each day's production. If we plot the 100 door weights recorded over time in the past 20 days, the resulting run chart appears as in Figure 9.4.

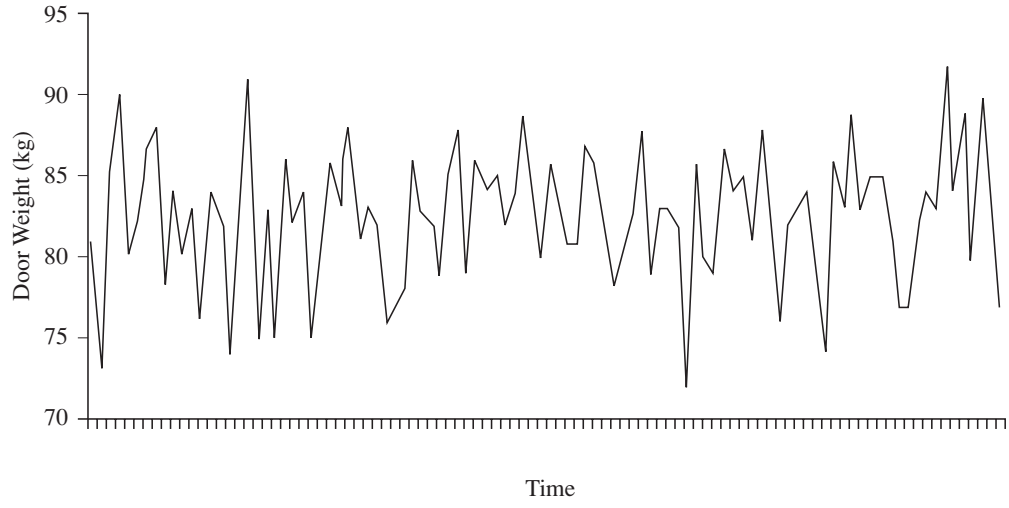


FIGURE 9.4 Run Chart of Door Weights Over Time

9.2.5 Multi-Vari Charts

To analyze the observed variability in process performance further, we may try to separate (1) variability *among* flow units produced at one time and (2) variability *between* flow units produced across time. Isolating the two types of variability would facilitate our search for and elimination of its source. To distinguish between the two types of variability, we take N samples of process performance over time, each sample containing n observations. For each sample, we then compute the highest, the lowest, and the average measurement.

A **multi-vari chart** is *a plot of high-average-low values of performance measurement sampled over time.* The range between high and low measurements within each sample indicates variability in flow units produced at one time, while fluctuations in sample averages show variability across time.

EXAMPLE 9.5

From the data in Table 9.1, we compute the high, low, and average weights $N = 20$ samples, each containing $n = 5$ observations and summarize the results in Table 9.2.

Table 9.2 Variation Between and Within Samples

	Day									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
High	90	88	84	91	86	88	83	86	88	89
Low	73	78	76	74	75	81	76	79	79	80
Average	81.8	83.8	81.0	80.8	80.4	83.8	79.2	83.0	84.4	83.8
	Day									
	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
High	87	86	88	87	88	84	89	85	92	90
Low	81	78	79	72	81	76	74	77	82	77
Average	83.8	82.0	83.0	80.8	83.8	80.8	83.0	81.2	85.0	83.8

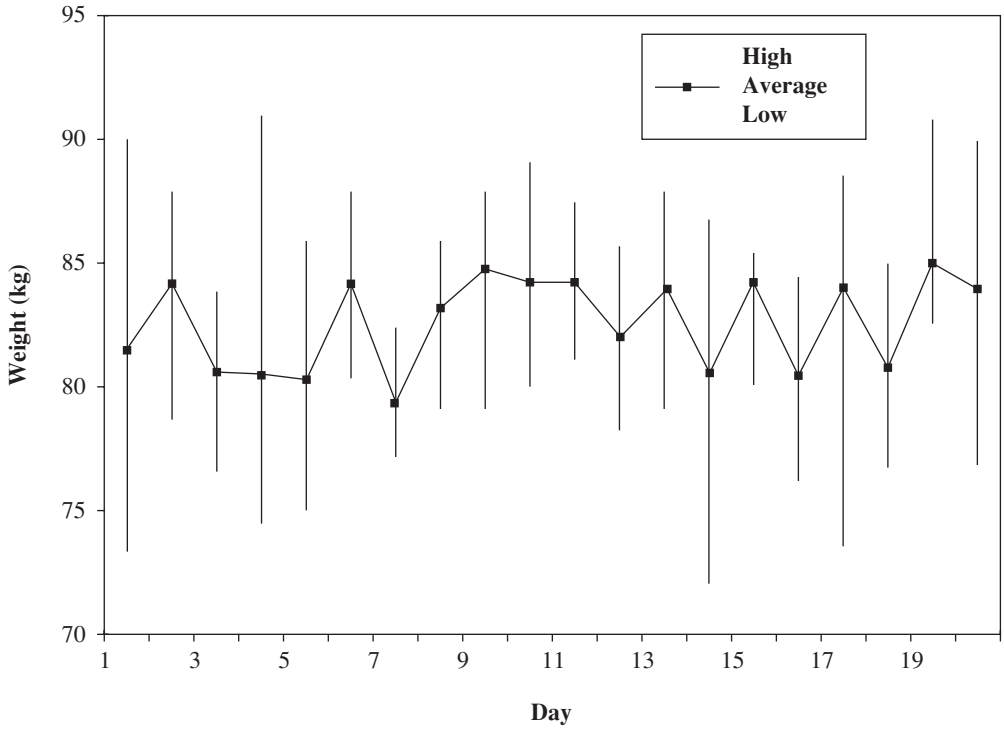


FIGURE 9.5 Multi-Vari Chart of Door Weight Variability

The high, low, and average values can now be plotted as a multi-vari chart in Figure 9.5. The length of each vertical line represents the range of variation in weights sampled on a given day, which indicates the amount of variability among the doors produced within that day's production. The middle dot on each vertical line shows the average weight of doors produced on that day. Fluctuation in the average weight from one day to the next then indicates variability between doors produced on different days, which is tracked by lines connecting the dots across time.

From this multi-vari chart, we see that there is relatively little fluctuation in average door weights across days. We may therefore conclude that there is no apparent trend or cyclical pattern over time that affects door weights between days (ruling out, e.g., "Friday afternoon" or "Monday morning" effects on worker performance). On the other hand, the lengths of vertical lines represent ranges of door weights produced within days, which seem to vary more from one day to the next. So, in our search to reduce overall variability, we should look for causes of variability that are common to all days rather than those that affect weights across days.

The basic idea of separating variability between and within batches is also useful in disaggregating variability between and within worker teams, work shifts, and so forth. The goal is to isolate different types of variability so that we can identify, control, and eliminate causes of the most prevalent type. However, beyond displaying variability within and across samples, multi-vari charts do not provide any guidance for taking actions.

From Figure 9.5, we note that on Day 19, the average door weight observed was 85 kg.—the highest of all averages observed so far and 3.8 kg. above the previous day's average. Should we have taken any action back on Day 19 to try to reduce the door weight? Well, we didn't, but luckily on Day 20 the average came down to 83.8 kg. In retrospect, it was good that we didn't overreact hastily. But what should we do if,

on Day 21, we get a sample with average weight of 86 kg.? Is that too high to be ignored? We need some operational decision rule for taking actions based on observed variability.

The same problem arises whenever one has to make decisions on the basis of new information as it becomes available over time. For example, an investor has to decide when to buy or sell a stock as its price fluctuates; the central bank has to decide when to raise or lower interest rates on the basis of the economic data collected over time. We devote the next section to the analysis of this important problem of deciding when to act and when not to act.

9.3 PROCESS CONTROL

As we saw in Chapter 1, there are two aspects to process management: process planning and process control. Process planning involves structuring the process, designing operating procedures and developing such key competencies as process capability, flexibility, capacity, and cost efficiency. In the long-run, process planning also involves process improvement aimed at producing and delivering products that will satisfy targeted customer needs better. The goal of process control, on the other hand, is to continually ensure that, in the short run, the actual process performance conforms to the planned performance. Now, the actual performance may deviate from the planned performance because of various disturbances. Process control involves tracking these deviations over time and taking corrective actions as deemed necessary.

9.3.1 The Feedback Control Principle

Central to managing process performance over time is the general principle of feedback control of dynamical systems, which involves two steps:

1. Collecting information about a critical performance metric over time
2. Taking corrective actions based on the observed performance to steer and maintain it at a desired level

Figure 9.6 displays the **feedback control principle**.

A house thermostat is a classic example of a feedback control mechanism. We set it at a desired temperature and a thermometer monitors the actual temperature, which may fluctuate because of air leaks, door openings, wind conditions, and so

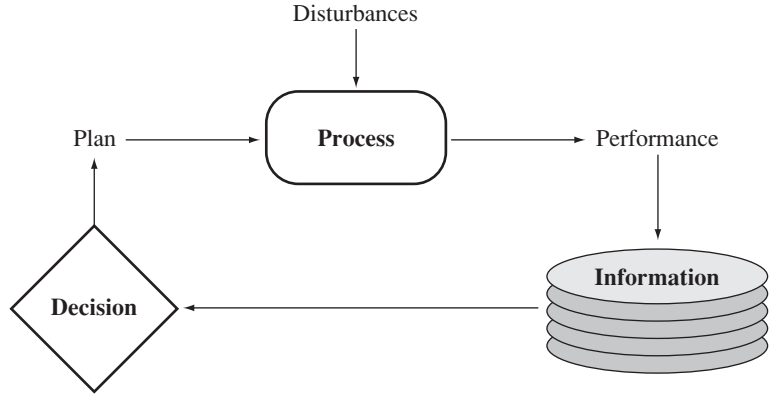


FIGURE 9.6 The Feedback Control Principle

forth. Depending on the actual temperature, the controller automatically turns the air conditioner or furnace on and off over time. Automobile cruise control is another example of a feedback control mechanism. It maintains the car speed by monitoring the actual speed and adjusting the fuel supply to ensure that the actual speed stays within a close range of the desired speed in spite of uneven hills encountered along the highway.

Applying the feedback control principle to process control involves periodically monitoring the actual performance (in terms of cost, quality, availability, and response time), comparing it to the planned levels of performance, identifying causes of the observed discrepancy between the two, and taking corrective actions to eliminate those causes.

Conceptually, process planning and control are similar to the **Plan-Do-Check-Act (PDCA) cycle** for problem solving and continuous improvement. The PDCA cycle *involves planning the process, operating it, inspecting its output, and adjusting it in light of the observed performance.* These four activities are then repeated continuously to monitor and improve the process performance over time.

The main challenge in process control is deciding *when* to act in light of the observed performance. In practice, process managers often compare the current period's performance with that of the previous (or a comparable) period in the past. Thus, cost and productivity variance reports typically show percentage gains or losses from one month to the next. Managers then base actions (e.g., granting rewards and reprimands) on whether the observed variances are favorable. Unfortunately, some variances may be due to factors beyond a subordinate's control, so any incentive mechanism based on such variance reports will be ineffective. According to the late quality guru W. Edwards Deming, incentives based on factors that are beyond a worker's control (which he called "system causes") is like rewarding or punishing workers according to a lottery or weather conditions. To base actions on the observed performance rationally, we must determine which variability in performance is due to factors that are within a subordinate's control and which are beyond his or her control. We must understand different types of performance variation and their causes, because appropriate managerial actions that are required to tackle each are very different.

9.3.2 Types and Causes of Variability

Every process displays variability. Some of this variability is *normal*—to be expected of the process of a given design, operating in a given environment—while *abnormal* variability also appears unexpectedly from time to time.

Normal variability is *statistically predictable and includes both structural variability and stochastic variability.* Recall that structural variability refers to systematic changes in the process performance, including seasonality and trend patterns. Stochastic variability refers to noise that arises due to *random* (*chance* or *common*) causes that are inherent to every process. Random causes are many in number, but each has only a small and unpredictable effect on the process performance. They cannot be isolated or removed easily without redesigning the entire process. For example, the weight of garage doors produced varies from door to door because of many factors bearing on the precision of the production process. A histogram of door weights shows the frequency distribution, while its average and standard deviation summarize the target door weight and the process precision in achieving it. Beyond that we cannot say why two consecutive doors from the same day's output have different weights; the production process is inherently imprecise. If the performance variability is normal, due to random causes only, the process is in a stable state of statistical equilibrium, that is, parameters of the probability distribution of its performance (e.g., the mean and the variance) are unchanging, the

process is performing as expected, given its design. How can we remove these random causes and increase consistency of our process performance? Only by improving the process design, which involves purchasing more precise equipment, hiring better skilled workers, training them well, purchasing better quality materials, and so forth. All this takes time and investment of resources over the long term and is therefore management's responsibility. It is unreasonable to expect line operators to produce consistent output when the process provided to them is inherently imprecise and the operating environment is unstable. Although line operators are invaluable in suggesting process improvements, the onus of these changes is on top management because of the scope, resources, and the time frame involved.

In contrast, **abnormal variability** *disturbs the state of statistical equilibrium of the process by changing parameters of its distribution in an unexpected way.* Abnormal variability results from *assignable* (or *special*) causes that are externally superimposed from time to time. The existence of abnormal variability means that one or more of the factors affecting the process performance—its architecture, procedures, or environment—may have changed. Although assignable causes are few in number, each has a significant effect on process performance. On the upside, however, they can be isolated, investigated, and eliminated, even in the short run. A particular batch of raw material might be defective, the machine may be incorrectly set, or the operator may be ill on that day. Because such causes are identifiable and correctable in the short run, at the local level, and without large capital expenditures, they can be delegated as the line operator's responsibility.

The goal of process control is to identify whether the observed variability in performance is normal or abnormal, so that an appropriate action can be taken to eliminate it.

Ironically, another source of abnormal variability arises from tampering with the process—making unnecessary adjustments in trying to compensate for normal variability. Deming's "marble experiment" illustrates this problem beautifully. A subject is asked to drop a marble through a funnel repeatedly with the goal of hitting a target on the floor underneath. If the marble misses the target, a naive subject tries to compensate for the deviation by moving the funnel in the opposite direction. This unnecessary tinkering, however, results in an *increase* in the variability of the marble's final position. The correct strategy, of course, is to aim the funnel right on the target and let the marble land around it, its final position exhibiting stochastic variability due to random causes. The idea is to avoid overreacting to random fluctuations in the short run. In the long run, we may wish to reduce even random fluctuations by redesigning the process—for example, by lowering the funnel, using a less bouncy marble, or leveling and changing the composition of the landing surface.

In statistical terms, normal variability is observed among random draws from a fixed probability distribution of process performance. Abnormal variability occurs when the parameters of this distribution (e.g., its mean or variance) are changed. Thus, in the short run, our goal is fourfold:

1. To estimate normal stochastic variability, separated from structural variability
2. To accept it as an inevitable part of the given process due to random causes and avoid unnecessary tampering to counteract it
3. To detect the presence of abnormal variability
4. To identify and eliminate assignable causes of abnormal variability

In the following sections, we assume that structural variability has already been accounted for and that tampering is avoided. As we monitor process performance over time, we wish to determine whether the observed performance variability is normal or abnormal. If it is normal—due to random causes only—we say that the process is in control. We should then accept the observed variability in performance as to be

expected and that it cannot be eliminated in the short run. It represents the best effort of the process, and we should leave it alone. If, on the other hand, performance variability is exceptional or abnormal—due to an assignable cause—we conclude that the process is out of control. In that case, we should stop the process and investigate, identify, and remove assignable causes, so as to bring it back into the state of control. The fundamental problem, therefore, is how to decide whether observed performance variability is normal or abnormal, that is, whether the process is in control or out of control.

9.3.3 Control Limit Policy

The basic idea is simply that if the process performance varies "too much" from the expected level, we should conclude that this variability is most likely abnormal, the process is out of control, and we should look for an assignable cause; otherwise, we should accept the observed variability as normal, the process is in control and no action is warranted. To quantify what we mean by variability being "too much," we establish a control band, which is a range within which any variation in process performance should be interpreted as normal, due to known structural or random causes that cannot be readily identified or eliminated in the short run. Therefore, if the process performance varies within this band, we should consider it as to be expected of the given process, and not tamper with it. Any variability outside this range, on the other hand, should be considered abnormal, due to some assignable cause, warranting a detailed investigation, identification, and correction.

This type of "threshold" policy for making decisions based on the observed performance has an intuitive appeal, and is known to be optimal in a wide variety of situations. As an everyday example, we might monitor the performance of our car by tracking the gas mileage we get from one fill-up to the next. If we get 25 miles per gallon on average, a combination of random causes (such as weather and traffic conditions) and assignable causes (out of tune engine or worn out tires) will make actual mileage to deviate from the expected one. Our decision rule may be to set a lower limit of acceptable mileage (say, 20 mpg). If the actual mileage falls below this limit, we should take the car to a mechanic for a checkup; if it is above this limit, we should continue to drive it as is. Similarly, in the house thermostat example, the temperature controller may be set at 20°C, and may turn the furnace on if the temperature drops 2°C below the set temperature and shut it off if the temperature rises 2°C above the set value. As a result, the house temperature will be maintained within 18°C and 22°C. A more precise thermostat may be more expensive to purchase but would maintain the room temperature closer to the desired setting by turning the furnace on and off more frequently.

Although the concept of process control is usually applied to managing product quality, the control limit principle is equally applicable to controlling any measure of process performance over time. For example, we have already seen application of a control limit policy in managing inventory and capacity. In Chapter 7, we saw that inventory control with uncertain demand involves monitoring the inventory level over time and ordering Q units as soon as the inventory depletes to a preestablished reorder point (ROP) level. In this context, the ROP constitutes a "control limit," and the action taken, when necessary, consists of ordering Q units. The ROP determines the safety inventory, which ensures product availability by limiting the probability of stockout. With periodic review, we set an upper limit U and a lower limit L, so the control limit policy is to order up to U if the inventory at review time is below L.

Similarly, in managing process capacity to limit waiting time, as we studied in Chapter 8, we may monitor the length of the waiting line (or the duration of a customer's waiting time). As soon as it reaches a specified upper limit U, we may increase

the process capacity by adding a server, and when it reaches a lower limit L, we may decrease the capacity. Such operating policies are routinely followed in opening and closing checkout counters in supermarkets, fast food restaurants and banks. In establishing the queue control limits U and L, the goal is to limit the customer's waiting time most economically. Similarly, in some service systems such as outdoor events, call centers, even hospital emergency rooms and ICUs, admission control policies turn away new arrivals if system congestion exceeds certain levels.

In the area of cost management, accountants use cost and productivity variance reports to track a department's performance and to specify taking managerial actions when the observed cost exceeds a certain threshold or productivity drops below a critical level. In short-term cash management, a firm's (or an investor's) cash position might fluctuate over time. If it falls below a certain level L, the firm may liquidate some of its assets in order to raise cash, while if the cash position reaches some higher level U, the firm may invest the excess cash in a riskier asset. Finally, in stock trading, investors can place "limit orders" to purchase (or "stop loss" orders to sell) a stock if and when its price drops to a specific level. Computerized "program trading" automatically executes trades when prices reach prespecified trigger levels.

Thus, in a wide variety of business applications, a control limit policy provides guidelines in the form of critical thresholds for taking actions online, in real time, in light of current information.

9.3.4 Control Charts

Statistical process control (SPC) involves establishing a control band of acceptable variation in process performance, comparing the actual performance against it through time and signaling when a corrective action is warranted. In setting the control band of acceptable variation around the mean μ, we should take into account two factors:

1. The normal variability in process performance, as measured by its standard deviation σ
2. How tightly we wish to control the process, as represented by a positive number z; the smaller the value of z, the tighter the control desired

We then set z standard deviations around the mean as the band of acceptable performance variation. We thus specify a **lower control limit (*LCL*)** and an **upper control limit (*UCL*)** and denote the **control band** as [*LCL*, *UCL*]. The general formulas for determining the two **control limits**, therefore, are

$$LCL = \mu - z\sigma \quad \text{and} \quad UCL = \mu + z\sigma \qquad \textbf{(Equation 9.1)}$$

Figure 9.7 shows a generic **control chart**, which *displays how process performance varies over time in relation to the control limits.* It is like a run chart of process performance, but with control limits overlaid to give it decision-making power to determine when to act and when not to act. As long as the observed performance varies within the control limits, we conclude that the variation is normal, due to random causes only, the process is in control, so no action is warranted. Any variation outside the control limits is to be regarded as abnormal, signaling an out-of-control process and probable presence of an assignable cause that should be investigated and eliminated.

In addition to comparing performance with control limits, one may also use additional rules for deciding when to act. For example, one rule recommends that if seven consecutive observations are above (or below) the average performance level, we should stop and investigate the process even though the variation is within the control band, because the probability of finding seven observations above the average is $(0.5)^7$ which is very small. Therefore, we should act in the interest of preventive maintenance.

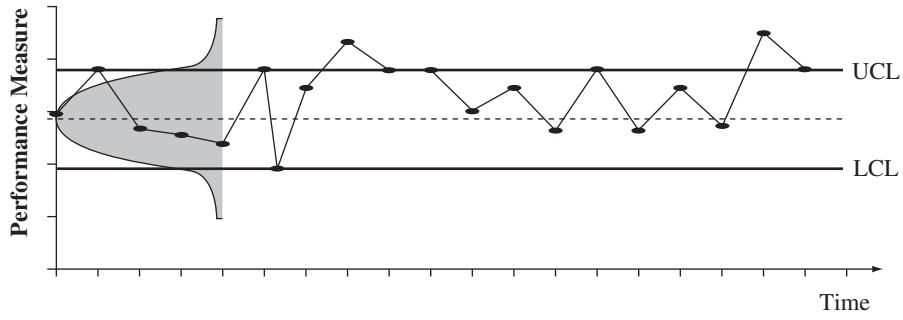


FIGURE 9.7 Process Control Chart

In addition to signaling the presence of an assignable cause, control charts also help us identify any structural variability in terms of trend or seasonal patterns over time and use this information to make decisions. They represent an outstanding example of graphical tools that are useful for monitoring and managing performance of any process over time.

Statistical Interpretation A reader familiar with statistics may recognize the relationship of control limits to hypothesis testing. We start with a "null hypothesis" that the process is in control (i.e., stable) at some level μ, the "alternate hypothesis" being that the process mean has in fact shifted to some other level. Based on the observed performance, we must determine whether to accept or reject the alternate hypothesis. The decision rule is to reject it and take no corrective action if the observed performance falls within the control limits; the evidence is not strong enough to support the alternate hypothesis that the process mean has shifted. If the performance measurement falls outside the control limits, we conclude that there is statistically significant evidence to reject the null hypothesis and accept the alternate hypothesis; the mean seems to have shifted, so we should look for an assignable cause.

This decision rule is not mistake-proof because normal variability may sometimes be misinterpreted as abnormal, and vice versa, leading to wrong decisions. Even when the process is in control with a stable mean, its performance measure may fall outside the control band simply because of normal variability. In that case, we may conclude—wrongly—that the process is out of control and look for an assignable cause when in fact none exists, leading to an expensive wild-goose chase. *The probability of false alarm due to mistaking normal variability as abnormal* is called **type I (or α) error**. Conversely, the process performance measure may fall within the control band just by chance, even if there is an assignable cause that shifts the mean. In this case, we conclude—again wrongly—that the observed variability is normal and warrants no investigation when in fact it is abnormal and we should be looking for its assignable cause. *The probability of missed signal due to mistaking abnormal variability as normal is called* **type II (or β) error**. In our car gas mileage example, suppose we usually get an average of 25 mpg and we set a lower control limit at 20 mpg, so that we take the car to a mechanic whenever the mileage drops below 20 mpg. It is possible that even when nothing is wrong with the car, our mileage may drop below 20 mpg purely by chance because of environmental and driving conditions; so we wrongly conclude that the car needs repair, which of course costs us unnecessary time, effort, and money. On the other hand, sometimes the car may in fact need a tune up and yet give an acceptable mileage above 20 mpg, leading us to ignore a problem that should be corrected. Thus, a decision rule based on a

control limit policy—although quite plausible—may lead to wrong conclusions, simply due to stochastic variability.

Optimal Degree of Control A key managerial challenge and discretion lies in choosing the degree of control exercised, which depends on two factors:

1. How frequently we monitor the process performance
2. How much variability in process performance we consider acceptable

Both of these factors affect the cost of control and the cost of operating a process out of control, and the challenge is to strike an optimal balance.

In automatic control systems (e.g., a house thermostat and car cruise control) monitoring and adjustments are performed continuously. In business processes, however, continuous control may not be economical or even possible. Frequent monitoring increases the cost of observation, although it improves the chance of quickly discovering any degradation in performance, leading to speedy recovery. The optimal frequency of monitoring should balance these costs and benefits. For example, the heart rate and cholesterol levels of a person with heart disease should be monitored more frequently than those of a healthy person, because the cost of monitoring is insignificant in relation to the risk of a heart attack.

Responsiveness of process control also depends on the width of the control band, as measured by z, which determines the magnitude of type I and type II errors and the resulting costs of making wrong decisions. From Equation 9.1, note that a smaller value of z means a narrower control band, which leads us to look for assignable causes more often than we should, resulting in frequent unnecessary investigation (or type I error). Investigation involves costly effort as well as lost output if the process must be stopped in the meanwhile. At the same time, however, the tighter control band ensures that assignable causes, when present, will be detected faster, which would reduce the cost of nonconformance (or type II error). Conversely, a larger z—and a wider control band—means looser control, infrequent investigation, and a lower cost of control but also a higher cost of nonconformance. The correct choice of z would balance these costs of investigation and failure to identify and eliminate assignable causes of variability. On the one hand, we would like to avoid a knee-jerk reaction to normal variability. On the other hand, we would like to discover and act promptly to eliminate any abnormal variability.

In summary, the optimum degree of control—in terms of the frequency of monitoring and the sensitivity of the decision rule—is based on tradeoffs between the costs of control and costs of nonconformance, as displayed in Figure 9.8. The long-run managerial

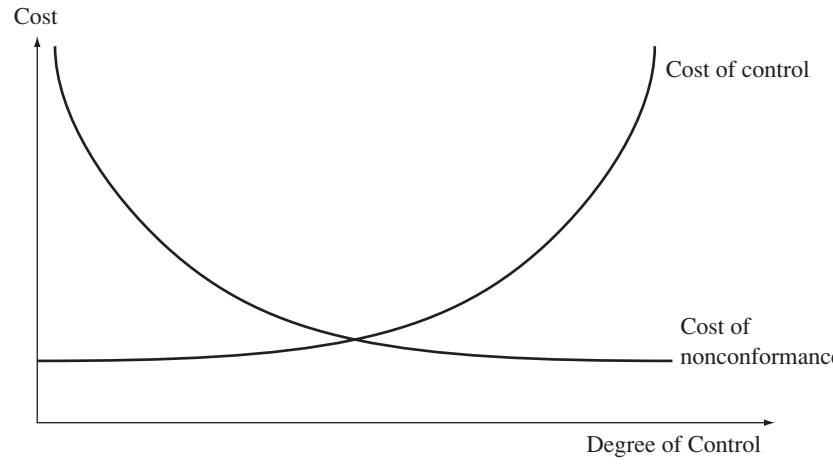


FIGURE 9.8 Optimal Degree of Control

challenge is to reduce the cost of control while setting tighter control limits. Toyota accomplishes this by hiring, training, and promoting smart team managers who can detect problems faster.

Although ideally, the optimal value of z should balance the costs of type I and type II errors involved, traditionally in the practice of statistical process control, a value of $z = 3$ is often used. Recall that if a performance measure is normally distributed, 99.73% of all measurements will fall within the mean ± 3 standard deviations, so $z = 3$ corresponds to type I error of 0.27%. Also in practical application of process control, we often do not know if a performance measurement is normally distributed, nor do we know its true mean or standard deviation. We must, therefore, ascertain these by sampling the actual performance and establish control limits based on sample estimates, as discussed in the next section.

Average (or $\overline{X}$) and Range (or R) Control Charts We monitor process performance by taking random samples over time. As we saw in Section 9.2.5, multi-vari charts display performance variability *within* each sample and *between* samples, but they do not tell us whether the observed variability is normal and hence should be left alone or is abnormal that warrants an action. Average and range charts accomplish this by establishing bands of acceptable variability in averages across samples and ranges within samples.

Suppose, as before, we take N random samples of process performance over time, each containing n observations. We compute two summary statistics for each sample:

- Sample average $\overline{X}$ of the n measurements
- Sample range R, which is the difference between the highest and the lowest measurements among n observations

Thus, we obtain N sample averages $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_N$ and ranges $R_1, R_2, \ldots, R_N$. Each sample average is an estimate of the expected (or mean) performance of the process, whereas sample range indicates variability in process performance (which is directly related to—but not the same as—the standard deviation and is easier to compute). As in a multi-vari chart, we can plot these sample averages and ranges over time, thereby displaying variability *between* and *within* samples, respectively. To decide whether observed variability is normal or abnormal we need to establish control limits on sample averages and ranges.

The average (or $\overline{X}$) control chart shows the control band of acceptable variability in averages across time, with the goal of identifying abnormal variability that affects the process mean. An important result in probability theory known as the central limit theorem states that the probability distribution of randomly taken sample averages will be approximately normal, even if individual observations are not. Therefore, we can assume that sample average $\overline{X}$ is normally distributed with some mean $\mu_{\overline{X}}$ and some standard deviation $\sigma_{\overline{X}}$. It turns out that $\mu_{\overline{X}} = \mu$, which is the same as the mean of each individual observation and $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, which is smaller than the standard deviation of the individual observation (that is, sample averages display less variability than individual observations). We can, therefore, apply the generic control limits of Equation 9.1 to obtain

$$LCL = \mu - z\sigma/\sqrt{n} \quad \text{and} \quad UCL = \mu + z\sigma/\sqrt{n}$$

where μ and σ are the true mean and the true standard deviation of the individual observations, both of which are typically unknown. We therefore estimate μ by the overall average $\overline{\overline{X}} = (\overline{X}_1 + \overline{X}_2 + \cdots + \overline{X}_N)/N$, and σ by s, the standard deviation of all Nn observations. With these estimates, we can obtain the control limits for the average ($\overline{X}$) control chart as

$$LCL = \overline{\overline{X}} - zs/\sqrt{n} \quad \text{and} \quad UCL = \overline{\overline{X}} + zs/\sqrt{n} \qquad \textbf{(Equation 9.2)}$$

We monitor sample averages over time. If they fall within the control band, we say that "the average (or $\overline{X}$) chart is in control" and conclude that the process mean is stable, the observed variability between sample averages must be due to normal causes only. If a sample average falls outside the control band, we conclude that the observed variability is abnormal, and we should look for an assignable cause that may have changed the process mean.

In addition to controlling the process mean, we would also like to make sure that variability in process performance is stable over time. As indicated before, a greater variability means a wider range of variation R in the observed performance within each sample. Given the observed ranges $R_1, R_2, \ldots, R_N$ in N samples, we can compute the average range $\overline{R} = (R_1 + R_2 + \cdots + R_N)/N$, which is a measure of performance variability, and s_R, the standard deviation of the sample range. With these estimates of the expected value and variability in sample variations, we apply the generic control limits to sample variations to obtain the range (or R) control chart as

$$LCL = \overline{R} - zs_R \quad \text{and} \quad UCL = \overline{R} + zs_R \qquad \textbf{(Equation 9.3)}$$

If observed ranges fall within this control band, we say that the "range (or R) chart is in control" and conclude that the process variability is stable, so that any observed variability within samples must be due to normal causes only. If an observed sample range is above the upper control limit, we should look for an assignable cause for excessive variability. If the observed range is below the lower control limit, the process performance is significantly better (more consistent) than we expected, which is a good sign; we should then try to find the reason for the change, reward it, and try to institutionalize it.

We illustrate construction of the $\overline{X}$ and R control charts for the data in Example 9.5.

EXAMPLE 9.6

As shown in Table 9.1, we have taken $N = 20$ samples of door weights over time, each containing $n = 5$ doors. As in Table 9.2, we can compute the average door weight in the sample of five on Day 1 as

$$\overline{X}_1 = (81 + 73 + 85 + 90 + 80)/5 = 81.8 \text{ kg.}$$

and the range between the heaviest and the lightest door in that sample as

$$R_1 = 90 - 73 = 17 \text{ kg.}$$

Similarly, the average weight on Day 2 is $\overline{X}_2 = 83.8$ kg. with the range of variation $R_2 = 10$ kg. and so on. These sample averages and ranges are tabulated in Table 9.3.

Table 9.3 Sample Averages and Ranges of Door Weights over Time

Sample	1	2	3	4	5	6	7	8	9	10
$\overline{X}$	81.8	83.8	81.0	80.8	80.4	83.8	79.2	83.0	84.4	83.8
R	17	10	8	17	11	7	7	7	9	9
Sample	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
$\overline{X}$	83.8	82.0	83.0	80.8	83.8	80.8	83.0	81.2	85.0	83.8
R	6	8	9	15	7	8	15	8	10	13

Note that both the averages and ranges in weights vary from one sample to the next. With the twenty averages from Table 9.3, we can now compute the grand average weight over all samples as

$$\overline{\overline{X}} = 82.5 \text{ kg}.$$

which, of course, matches the overall average weight of all 100 doors sampled, as calculated in Example 9.3, where the standard deviation of the individual door weights was calculated to be $s = 4.2$ kg.

If we accept the standard practice of setting $z = 3$, our control limits on sample averages in Equation 9.2 become

$$UCL = \overline{\overline{X}} + z\,s/\sqrt{n} = 82.5 + (3)(4.2)/\sqrt{5} = 88.13 \text{ kg}.$$
$$LCL = \overline{\overline{X}} + z\,s/\sqrt{n} = 82.5 - (3)(4.2)/\sqrt{5} = 76.87 \text{ kg}$$

If we compare the 20 values of sample averages in Table 9.2 against these limits, we see that all of them fall within our control band [76.87, 88.13]. Equivalently, we can plot the upper and lower control limits on the chart in Figure 9.9, and see that all points fall within the control limits. So we conclude that the process mean is stable; there is no statistical evidence to indicate the presence of an assignable cause of variability that affects the process mean. In other words, there is no reason to believe that door weights vary significantly between days.

Likewise, we can compute from Table 9.3 the average range

$$\overline{R} = 10.1 \text{ kg}.$$

and standard deviation of ranges

$$s_R = 3.5 \text{ kg}.$$

and establish control limits on values of the observed ranges as

$$UCL = \overline{R} + z\,s_R = 10.1 + (3)(3.5) = 20.6$$
$$LCL = \overline{R} - z\,s_R = 10.1 - (3)(3.5) = -0.4$$

Note that we should set the LCL of -0.4 to 0 because the range of variation within a sample cannot be negative. Again, when we compare the 20 observed ranges against these control limits, we see that they are all less than 20.6 kg. Equivalently, we can plot the observed ranges against the upper and lower control limits as shown in Figure 9.10.

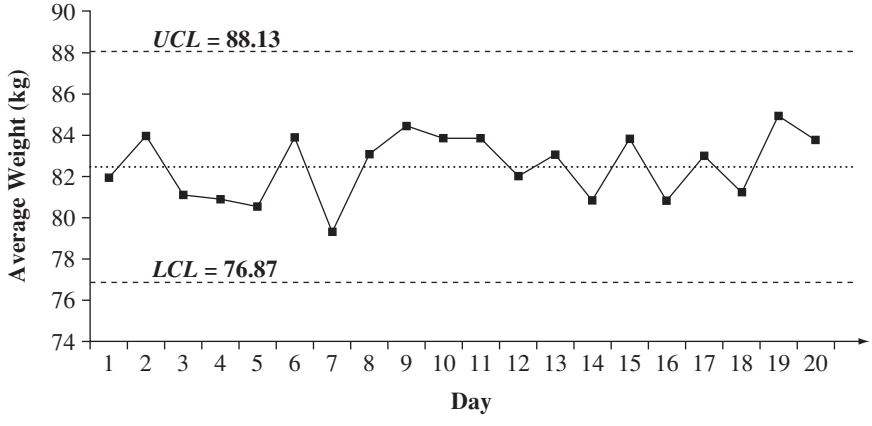


FIGURE 9.9 Average ($\overline{X}$) Control Chart

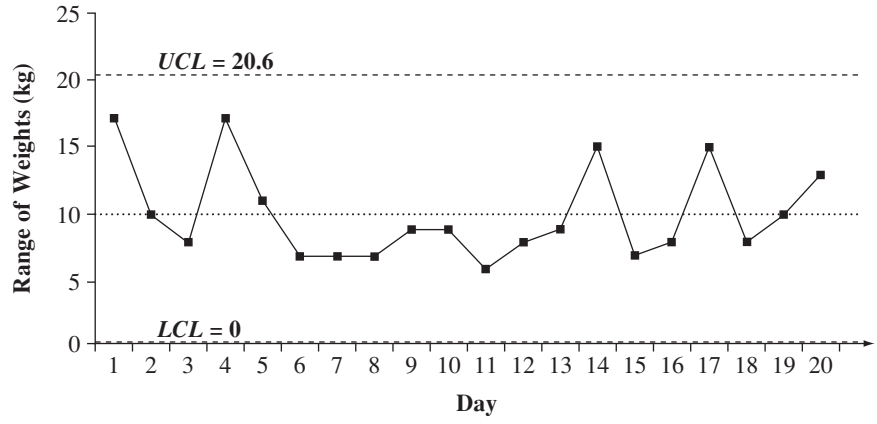


FIGURE 9.10 Range (*R*) Control Chart

Observe that all points are below $UCL = 20.6$. We, therefore, conclude that there is no reason to believe that any day's output is varying significantly either. In other words, there is no assignable cause of variability within each day's performance.

Thus, we conclude that our production process seems to display only normal variability within as well as between days. In other words, the process is in control; and as far as door weights are concerned, the door-making process appears to be statistically stable.

In order to highlight the essence of statistical process control, we have described the terminology and technical details of control charts in a somewhat simplified form. For example, we have used sample standard deviation s as an estimate of the true standard deviation σ, and standard deviation of sample ranges s_R as an estimate of the true standard deviation of ranges σ_R. In conventional statistical process control, both σ and σ_R are estimated by constant multiples of the average range $\overline{R}$, where the values of these constants depend upon the sample size n. Details may be found, for example, in Grant and Leavenworth (1988). Note that both the X-bar and R charts are required to track a process accurately.

Although we have described process control in terms of the quality of its output, the same principles would apply if we wish to control the process with respect to other performance measures such as the unit flow time or cost. Since these metrics can be measured on a continuous scale, they may be assumed to be normally distributed. Sometimes, performance may be measured in terms of a discrete variable, such as number of defective units produced, or number of defects found per flow unit. In such cases, we need to use an appropriate discrete probability distribution to derive the control limits, but the basic principles of a control chart remain the same, as will be evident from the following section.

Fraction Defective (or *p*) Chart Instead of a detailed measurement of a quality metric, we may choose to classify each flow unit as "defective" or "nondefective" based on its overall quality in meeting customer requirements. If the process is producing fraction defective p, and if we take a random sample of n flow units, then the number of defectives D in the sample will have binomial distribution with parameters n and p, which has mean np and variance $np(1 - p)$; see Appendix II. The fraction defective $P = D/n$ will then have mean p and variance $p(1 - p)/n$. To estimate the true fraction defective p, we take N samples, each containing n flow units, observe proportion defective in each and compute the average fraction defective $\overline{p}$. The **fraction defective (or *p*) chart** *shows control limits on the observed fraction of defective units as*

$$LCL = \overline{p} - z\sqrt{\overline{p}(1 - \overline{p})/n} \text{ and } UCL = \overline{p} + z\sqrt{\overline{p}(1 - \overline{p})/n} \qquad \textbf{(Equation 9.4)}$$

To illustrate, suppose in our garage door example, we classify each door simply as defective or good, depending on its overall quality such as fit and finish, dimensions, weight, etc. Based on 20 samples of 5 doors each, suppose we find the number of defective door in each sample batch to be 1, 0, 0, 2, 1, 1, 0, 1, 2, 1, 2, 1, 1, 2, 1, 0, 3, 0, 1, and 0. Dividing each by 5 gives fraction defective in each sample as 0.2, 0, 0, 0.4, 0.2, 0.2, 0, 0.2, 0.4, 0.2, 0.4, 0.2, 0.2, 0.4, 0.2, 0, 0.6, 0, 0.2, and 0. The average proportion defective is then $\bar{p} = 0.2$. With $z = 3$, the control limits on the fraction defective become

$$UCL = 0.2 + (3)\sqrt{0.2(1 - 0.2)/5} = 0.7366$$
$$LCL = 0.2 - (3)\sqrt{0.2(1 - 0.2)/5} = -0.3366$$

which should be rounded up to 0 since fraction defective cannot be negative. Thus, if the observed fraction defective is less than 0.7366, we conclude the process is in control, as is the case above.

Number of Defects (or *c*) Chart Suppose we wish to control the number of defects on each flow unit. Suppose n is the number of opportunities for a defect to occur and p is the probability that each actually materializes. Then, as before, the number of defects per flow unit will have binomial distribution with parameters (n, p). However, if n is large and p is small, the binomial distribution may be approximated by the Poisson distribution with mean $c = np$, which is also its variance; see Appendix II. We can estimate the true mean c by $\bar{c}$, the average number of defects per unit observed by sampling. We can then set **number of defects (or *c*) chart** *that shows control limits on the observed number of defects per flow unit as*

$$LCL = \bar{c} - z\sqrt{\bar{c}} \quad \text{and} \quad UCL = \bar{c} + z\sqrt{\bar{c}} \qquad \textbf{(Equation 9.5)}$$

If the observed number of errors exceeds the upper control limit, it indicates statistically significant degradation in performance that should be investigated. Similarly, if the observed number of defects is less than the lower control limit, it indicates better-than-expected performance that should be recognized and rewarded. In either case, whenever we get a signal that performance variability is abnormal, we should look for an assignable cause—favorable or unfavorable—and act on it.

To illustrate, suppose we wish to monitor and control the number of order processing errors that occur per month at Overhead Door. If they process several orders per month and the chance of making an error on each order is small, then the number of errors per month follows a Poisson distribution. Suppose they have tracked order processing errors over the past 12 months and found them to be 3, 1, 0, 4, 6, 2, 1, 2, 0, 1, 3, and 2. Then the average number of errors per month is $\bar{c} = 2.083$, so control limits are:

$$UCL = 2.083 + (3)\sqrt{2.083} = 6.413$$
$$LCL = 2.083 - (3)\sqrt{2.083} = -2.247$$

which should be rounded up to 0, since number of errors cannot be negative. Since all observed processing errors are less than 6.413 (even though we made 6 order processing errors in month 5), we conclude that the order processing process is in control.

Dynamics of Process Control Charts Note that in order to establish control limits as mean $\pm z$ standard deviation, we first need to estimate the true mean and the true standard deviation of the performance measure using sampled observations. To ensure that our estimates are reliable, it is essential that samples are randomly selected and drawn from a stable probability distribution with constant mean and standard deviation, which means the sample should be from the output of a process that is already in control. Thus, the logic of control charts may appear somewhat circular: Control limits are based on estimates of process parameters assuming that the process is in control but

then the same observations are compared against these control limits to determine whether the observed variability is normal! The apparent contradiction disappears if we view process control as an on-going activity which uses control charts as a dynamic tool for continually monitoring and estimating process performance. We compare the observed performance against the currently established control limits which are based on the estimates obtained from the data observed thus far. We then take action to stabilize the process if the current control limits are exceeded. As we continue to observe and stabilize the process performance over time, we keep improving our estimates of the parameters of the probability distribution of the performance measure. We adjust the control limits accordingly, and compare newly observed performance against these more reliable limits, and continue monitoring.

What do we do if process performance falls outside control limits? How do we investigate and correct causes of abnormal variability? We indicate two tools for systematically analyzing and correcting sources of abnormal variability.

9.3.5 Cause–Effect Diagrams

On detecting the presence of abnormal variability, we may use a **cause–effect diagram** (also known as a **fishbone diagram** or **Ishikawa diagram**) to identify the root cause(s) of the observed variability. A cause–effect diagram *shows a chain of cause–effect relationships that ultimately leads to the observed variability.* Through group discussion and brainstorming, we first try to generate hypotheses about possible causes. According to one guideline, if we diligently pursue a sequence of five *why?* questions, we will ultimately arrive at the root cause of a problem. For example:

- Why are these doors so heavy? *Because the sheet metal used to make them was too thick.*
- Why was the sheet metal too thick? *Because the rollers at the supplier's steel mill were set incorrectly.*
- Why were the supplier's rollers incorrectly set? *Because the supplier does not have expertise to produce to our specifications.*
- Why did we select a supplier who can't meet our specifications? *Because our project supervisor was too busy "getting the product out" to invest sufficient time in participating in vendor selection.*
- Why did he find himself in these circumstances? *Because he gets paid by his performance in meeting production quotas.*

Thus, the root cause of the door weight problem boils down to the company's incentive structure. A simplified fishbone diagram of this problem may look like the one in Figure 9.11. The tail of each arrow shows a possible cause of the effect indicated at the head of that arrow.

Although we have used a simplified example for illustration, in the real world, root cause determination is often a nontrivial problem. For example, sudden acceleration experienced by some drivers of Toyota and Lexus cars in 2009 led to expensive recalls, law suites, bad publicity, and loss of market share for the automaker well known for high quality. The root cause analysis of the problem involved several months of expert investigation into possible causes, including misfitting floor mat, faulty accelerator pedal design, software glitch, even driver error. Finally, the expensive electronic problem was ruled out in February 2011. As products have become increasingly complex with sophisticated electronics involved, tracing the root cause(s) of quality problems has become ever more challenging.

As another example, the worst off-shore oil spill disaster in the U. S. history occurred in 2010 in deep water drilling by British Petroleum (BP). The oil rig explosion

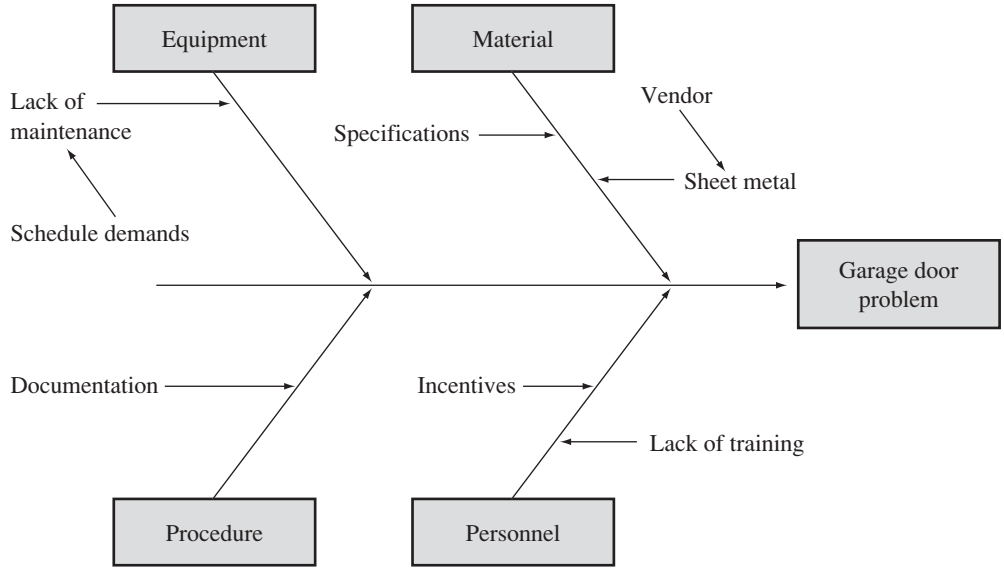


FIGURE 9.11 Cause-Effect Diagram of the Garage Door Problem

killed 11 workers and leaked 4.9 million barrels of oil into the Gulf of Mexico, resulting in severe damage to the environment (marine and wild life) and economy (fishing and tourism). A detailed investigation of causes involved several specific factors such as bad cement job and defective blowout preventer, as well as systemic failures by management of BP and its contractors in effective communication, training, and prevention measures. Finally, in March 2011, faulty design of the blowout preventer was identified as the cause of the disaster. Again, cause–effect analysis of a problem deep at the sea floor involved much more than drawing a simple fish bone diagram. However, it does provide a simple and systematic approach to problem solving.

Although a cause–effect diagram enables us to identify a *qualitative* relationship between a process variable and its effect on the product characteristic that customers care about, in order to take a concrete action, we need to understand the precise *quantitative* relationship between the two, as indicated in the next section.

9.3.6 Scatter Plots

Suppose we have identified the supplier's sheet metal rolling process as the root cause of the door-weight problem (which affects the ease of door operation and its durability that customers value). We would now like to measure the exact relationship between the two so that we will be able to control the door weight by changing the settings on the supplier's rolling mill.

To estimate this relationship, we may experiment with various settings on the rolling mill, measure the effect on the garage-door weights, and plot the results on a graph, which is called a **scatter plot**. Formally, a scatter plot is *a graph showing how a controllable process variable affects the resulting performance characteristic.* In the scatter plot shown in Figure 9.12, the horizontal axis represents the sheet metal thickness setting in millimeters, and the vertical axis shows the weight of garage doors produced. The two variables seem to be "positively correlated"—a higher roller settings tends to be associated with increased door weights, as one would expect. One could continue with statistical regression analysis to estimate the exact relationship, but we will not pursue it here. Suffice it to say for now that we have traced the root cause that affects door weights, which is critical in meeting customer requirements.

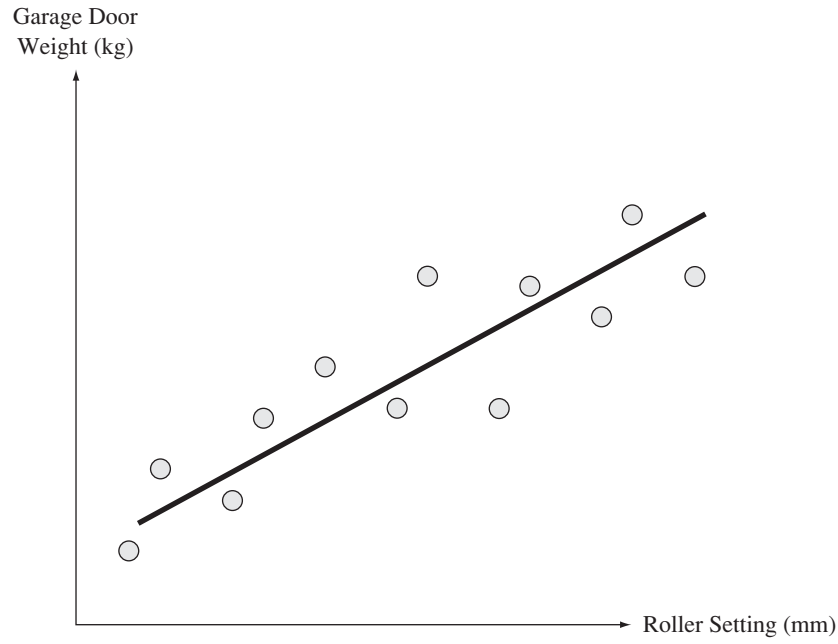


FIGURE 9.12 Scatter Plot

To summarize this section, process control involves dynamically monitoring process performance over time to ensure that performance variability is only due to normal random causes. It enables us to detect abnormal variability so that we can identify and eliminate its root causes. A process being "in control" means that variability in its performance is stable over time so that output is statistically predictable. Being in control makes a statement about the internal stability of the process. However, it is important to note that being in control does *not* necessarily mean that the process performance is satisfactory in terms of its output from the external customer's perspective! Therefore, beyond maintaining the process in a state of internal control, it is important for process managers to make sure that process performance also meets the external customer requirements—a topic that we take up in the next section.

9.4 PROCESS CAPABILITY

In our study of process planning and control, we first identified the external product measures that customers desire (e.g., the ease of door operation and durability) and linked them to internal measures (door weight) that the manufacturer can control. We then translated the product performance desired by customers into the necessary **upper specification (*US*)** and **lower specification (*LS*)** limits (e.g., 75–85 kg.) of the product design, which indicate *the range of performance variation that a customer is willing to accept.* Thus, product specifications represent performance variability that is acceptable from the external perspective of the customer. On the other hand, process control limits represent the range of performance variation that is acceptable from the internal perspective of maintaining process stability. Thus, it is important to note that process control limits and product specification limits serve very different roles and should not be mixed together, for example, by plotting both on the same chart. Thus in an average ($\overline{X}$) control chart, we plot and compare sample averages with control limits, and not with

specification limits. Similarly, individual units—not sample averages—must meet customer specifications.

Once our process is under control, so that its output is statistically predictable, our estimates of the process mean and standard deviation will be reliable. Based on these estimates, we can determine the **process capability**, which may be defined broadly as *the ability of the process to meet customer specifications.* Although we can measure process capability in a variety of ways, here we describe three that are closely interrelated.

9.4.1 Fraction of Output within Specifications

One measure of process capability is the fraction of the process output that meets customer specifications. We can compute this fraction either by actual observation or by using a theoretical probability distribution, as illustrated in Example 9.7.

EXAMPLE 9.7

Recall that in the garage door example, weight specifications are $LS = 75$ kg. to $US = 85$ kg. Also recall that in Figure 9.3, the height of each bar corresponds to the fraction of doors with a specific weight. Adding these bar heights between 75 and 85 kg., therefore, yields the total fraction of door output that meets design specifications. We see that 73 out of the 100 doors observed fall within the given specifications. We may, therefore, say that the process is currently 73% capable of meeting customer requirements; it is producing approximately 27% defectives.

Alternatively, we may use the normal distribution as a continuous approximation and compute the area under the normal probability density curve between 75 and 85 kg.

If door weight X is a normal random variable with mean $\mu = 82.5$ kg. and standard deviation of $\sigma = 4.2$ kg., then the proportion of doors falling within the specification limits is given by

$$\text{Prob}(75 \leq \mathbf{X} \leq 85) = \text{Prob}(\mathbf{X} \leq 85) - \text{Prob}(\mathbf{X} \leq 75)$$

Using Microsoft Excel, we get Prob $(\mathbf{X} \leq 85)$ = NORMDIST(85,82.5,4.2,True) = 0.724158, and Prob $(\mathbf{X} \leq 75)$ = NORMDIST(75,82.5,4.2,True) = 0.037073, so

$$\text{Prob}(75 \leq \mathbf{X} \leq 85) = 0.724158 - 0.037073 = 0.687085$$

therefore, the door-making process is capable of producing about 68.7% of doors within the specifications, or the company is delivering about 31.3% defective doors.

Note that, on average, doors weigh 82.5 kg., which is well within the specification limits, but that is not a relevant criterion for meeting customer requirements. Specifications refer to individual doors, *not* sample averages: We cannot comfort an individual customer by assuring that, on average, its doors do meet specifications, because there is a 31.3% chance that the customer will receive a door that is either too light or too heavy. It is the variability in individual doors—not just their average weight—that matters in determining how capable the process is in meeting customer requirements. As a quote goes, "Company may celebrate the average but customers are bothered by the variance."

What are the financial implications of defectives? A defective product or service results in cost of recall, rework, repair, and, ultimately, reputation. Unintended acceleration in some of Toyota's vehicles resulted in a recall of 8 million vehicles in 2009 at a cost of $15 million and a significant loss of market share to the competitors. Usually, early detection and correction of defectives cost much less than failure in

customer's hands in accordance with "a stitch in time saves nine," while prevention of defectives is even more effective and economical, hence the expression "quality is free." The conventional wisdom was that providing high quality requires more time and resources, whereas "doing it right the first time" may actually save more in the long run. Sometimes, even a minor defect may turn out to be catastrophic, as in an old proverb "for want of a nail, the horse shoe fell, toppled the warrior, and lost the war." In summary, defects are expensive and their prevention should be the goal of every process manager.

9.4.2 Process Capability Ratios (C_{pk} and C_p)

A related measure of process capability that is easier to compute is called the process capability ratio, denoted as C_{pk}. This measure is based on the observation that for a normal distribution, if the mean is 3 standard deviations above the lower specification LS (or below the upper specification US), there is very little chance of a product characteristic falling below LS (or above US). We therefore compute

$$(US - \mu)/3\sigma$$

and

$$(\mu - LS)/3\sigma$$

as surrogate measures of how well process output would fall within our specifications. The higher these values, the more capable the process is in meeting specifications. In fact, to be on the conservative side, we may take the smaller of these two ratios and define a single measure of process capability as

$$C_{pk} = \min[(US - \mu)/3\sigma, (\mu - LS)/3\sigma] \qquad \textbf{(Equation 9.6)}$$

A process with a higher value of C_{pk} is more capable than one with a lower value. Typically, a process with of C_{pk} of 1 or more represents a capable process that will produce most of the output that meets customer specifications.

The C_{pk} measure is also useful when our product specifications are one sided—that is, when we need to ensure that performance measurements are not too high (or too low). For example, if we need to measure the processing cost, delivery time, or number of errors per transaction, we may specify only the upper specification limit because lower values mean only better-than-expected quality. The C_{pk} is then given by the single term in the previous expression that is relevant, the first one in these examples.

As a special case, if the process is properly centered at the middle of the specification range, we may define C_{pk} by either

$$(US - \mu)/3\sigma$$

or

$$(\mu - LS)/3\sigma$$

as both are equal for a centered process. Therefore, for a correctly centered process, we may simply define the process capability ratio denoted by C_p as

$$C_p = (US - LS)/6\sigma \qquad \textbf{(Equation 9.7)}$$

This ratio has a nice interpretation. Its numerator specifies the range of performance variability that the customer is willing to tolerate (and so represents the "voice of the customer"). The denominator, meanwhile, denotes the range of variability that the process is capable of delivering (which represents the "voice of the process"). Recall

that with normal distribution, most process output—99.73%—falls within ±3 standard deviations from the mean. That is, most of the process variability is within 6 standard deviations around the mean. Consequently, 6σ is sometimes referred to as the "natural tolerance" of the process. We illustrate the computations of the process capability ratio for the door problem.

EXAMPLE 9.8

In our garage door example, since the mean is 82.5 kg. and the standard deviation is 4.2 kg., we can compute

$$\begin{aligned} C_{pk} &= \min[(US - \mu)/3\sigma\,,(\mu - LS)/3\sigma] \\ &= \min\{(85 - 82.5)/[(3)(4.2)]\,,(82.5 - 75)/[(3)(4.2)]\} \\ &= \min\{0.1984, 0.5952\} = 0.1984 \end{aligned}$$

If the process is correctly centered at $\mu = 80$ kg., we can compute the process capability ratio as

$$\begin{aligned} C_p &= (US - LS)/6\sigma \\ &= (85 - 75)/[(6)(4.2)] = 0.3968 \end{aligned}$$

It is important to note that $C_{pk} = 0.1984$ (or $C_p = 0.3968$) does *not* mean that the process is capable of meeting customer needs 19.84% (or 39.68%) of the time; we computed that figure in Example 9.7 to be about 69%.There is, however, a close relationship between the process capability ratio and the proportion of the process output that meets customer specifications, based on the standard deviation of performance variability. Table 9.4 summarizes this relationship, wherein defects are counted in parts per million (ppm) or parts per billion (ppb), and the process is assumed to be properly centered. Thus, if we would like no more than 100 defects per million (0.01% defectives), we should have the probability distribution of door weights so closely concentrated around the mean that the standard deviation is 1.282 kg., which then corresponds to $C_p = 1.3$.

9.4.3 Six-Sigma Quality

A third equivalent measure of process capability that has been employed by Motorola, General Electric, and other quality conscious companies is called the **sigma [capability] measure**, which is computed as

$$S = \min[(US - \mu)/\sigma\,,(\mu - LS)/\sigma] \qquad \textbf{(Equation 9.8)}$$

and the process is called an S-sigma process. If the process is correctly centered at the middle of specifications, Equation 9.7 is equivalent to

$$S = [(US - LS)/2\sigma] \qquad \textbf{(Equation 9.9)}$$

Table 9.4 Relationship between Process Capability Ratio and Proportion Defective

Defects (ppm)	10,000	3,000	1,000	100	10	1	2 ppb
C_p	0.86	1	1.1	1.3	1.47	1.63	2

EXAMPLE 9.9

Currently the sigma capability of the door making process is

$$S = \min[(85 - 82.5)/(4.2), (82.5 - 75)/4.2] = \min[0.5952, 1.7857] = 0.5952$$

By centering the process correctly, its sigma capability increases to

$$S = (85 - 75)/[(2)(4.2)] = 1.19$$

Thus, with a three-sigma process that is correctly centered, the upper and lower specifications are three standard deviations away from the mean, which corresponds to $C_p = 1$, and 99.73% of the output will meet the specifications. Similarly, a correctly centered **six-sigma process** has a standard deviation so small that the upper and lower specification limits are six standard deviations from the mean each. This level of performance consistency represents an extraordinarily high degree of precision. It corresponds to $C_p = 2$, or only two defective units per billion produced! In order for the door-making process to be a six-sigma process, its standard deviation must be

$$\sigma = (85 - 75)/[(2)(6)] = 0.833 \text{ kg.}$$

which is about one-fifth of its current value of 4.2 kg.

Adjusting for Mean Shifts Actually, given the sigma measure, Motorola computes the fraction defective more conservatively by allowing for a shift in the mean of ±1.5 standard deviations from the center of specifications. Allowing for this shift, a six-sigma process amounts to producing an average of 3.4 defective units per million produced. Thus, even if incorrectly centered, the six-sigma process will produce only 3.4 ppm defectives. Such a high standard represents, although not quite "zero defects," "virtual perfection" and a goal to strive for.

With these three measures of process capability and allowing for a 1.5-sigma shift in the process mean, we can determine the relationship between the sigma measure, C_p, and defective ppm produced, tabulated as in Table 9.5.

Why Six-Sigma? From the table, note that improvement in the process capability from a three-sigma to a four-sigma process calls for a 10-fold reduction in the fraction defective, while going from a four-sigma process to a five-sigma process requires a 30-fold improvement, and improving from a five-sigma to a six-sigma process means a 70-fold improvement. Thus, further improvements in process capability becomes increasingly more challenging. Experts estimate that an average company delivers about four-sigma quality, whereas best-in-class companies aim for six-sigma.

Why should we insist on such high—and perhaps unattainable—standards? For one thing, even if individual parts (or processing steps) are of extremely high quality, the overall quality of the entire product (or process) that requires *all* of them to work satisfactorily will be significantly lower. For example, if a product contains 100 parts and each part is 99% reliable, the chance that the product (all its parts) will work is only $(0.99)^{100} = 0.366$, or 36.6%!

Table 9.5 Fraction Defective and Sigma Measure

Sigma *S*	3	4	5	6
Capability Ratio C_p	1	1.33	1.667	2
Defects (ppm)	66810	6210	233	3.4

Moreover, even if defectives are infrequent, the cost associated with each may be extremely high. Deaths caused by faulty heart valves, automobile brake failures, or defective welds on airplane bodies, however infrequent, are too expensive to the manufacturers (in terms of lawsuits and lost reputation), customers (in terms of lives), and ultimately to the society. In fact, some consequences of product failures (such as BP's Gulf oil spill due to failure of the blowout preventer) may be immeasurable in terms of their impact on environment and human life.

Moreover, the competition and customer expectations keep rising constantly, and ambitious companies and their leaders continue to set such stretch goals such as six-sigma quality. Above all, pursuit of six-sigma quality represents a mindset and an organizational culture of continuous improvement in journey to perfection, however stretched the goal might be. This six-sigma philosophy is consistent with that of lean operations, which we will study in the next chapter. Lean operations involve eliminating waste of all kinds: Waste of material, defects, delays, space, movement, etc. The combination is often referred to as lean six-sigma process improvement.

Interestingly, airline baggage handling is only a four-sigma process (an American airline average is about 7 bags mishandled per 1000 flown), whereas their crashworthiness is better than a six-sigma process; thus a process *can* be perfected if one's life depended on it. Even in non-life-threatening contexts, AT&T's dial tone was available 99.999% of the time, which comes to availability of all but 5.26 minutes per year, as does Google's search service. Another example of a six-sigma process involves Mumbai's 5000 "dabbawalas," who pick up, deliver, and return tiffin lunch boxes from 200,000 homes and apartments to 80,000 office locations that are situated over 40 miles away, in three hours each way, without using any fuel or modern technology. The distribution process involves an ingenious combination of coding, aggregating, and sorting boxes and moving them in crates through public trains, push carts, even bicycles from each household to a correct office destination and back to home. The error rate of delivery is about 1 in 16 million trips, while the cost of service to customers is about $6 per month!

Safety Capability In general, we may also express process capability in terms of the design margin $[(US - LS) - z\sigma]$ and interpret it as safety capability, analogous to safety inventory, safety capacity, and safety time studied in Chapters 7 and 8. Each of these safety margins represents an allowance planned to meet customer requirements of product quality, availability, and response time in face of variability in supply and/or demand.

Greater process capability means less variability and less chance of failing to meet customer specifications. Moreover, if the process output is closely clustered around its mean, in relation to the width of customer specifications, most of the output will fall within the specifications, even if the mean is not centered exactly at the middle of the specifications. Higher capability, thus, means less chance of producing defectives even if the process goes out of control because of a shift in the mean off from the center of the specifications. Thus, process capability measures robustness of the process in meeting customer specifications. A robust process will produce satisfactory output even when it is out of control.

One criticism in measuring process capability in relation to customer specifications is that any performance within the given specification limits is considered equally acceptable. Genichi Taguchi's quality philosophy, on the other hand, suggests that being right on target is more important than just being within specifications. Even if all parts of a product (or steps in a process) are within specifications, its overall performance may not be satisfactory, because of "tolerance stacking," which means, for example, that two parts at the opposite ends of their specifications will not fit properly. Taguchi, therefore, suggests measuring loss in quality by the squared deviation in the actual performance from its target, as was mentioned in Section 9.1.

9.4.4 Capability and Control

As we saw in Example 9.7, based on observed data, the door production process is not performing well in terms of meeting the customer specifications; only about 69 percent of the output meets the specifications. Yet recall from Example 9.6 that we concluded that the door making process was "in control"! It is therefore important to emphasize that being in control and meeting specifications are two very different measures of process performance. Whereas being "in control" indicates *internal* stability and statistical predictability of the process performance, "meeting specifications" measures its ability to satisfy *external* customer requirements. Being in control is a necessary but not sufficient condition for satisfactory performance of a process. Measurements of a process in control ensure that the resulting estimates of the performance mean and standard deviation are reliable so that our assessment of the process capability is accurate. The next step is, then, to improve the process capability so that its output will be satisfactory from the customers' viewpoint as well.

9.5 PROCESS CAPABILITY IMPROVEMENT

Since each measure of process capability defined previously depends on both performance mean and standard deviation, we must try to adjust one or the other or both to improve the process capability.

9.5.1 Mean Shift

Given the probability distribution of process output, changing the process mean will shift the distribution and increase the proportion of output that falls within the specifications as well as the process capability ratio.

EXAMPLE 9.10

Clearly, the average door weight of 82.5 kg. is too high in relation to the customer specification of 75 to 85 kg. The histogram in Figure 9.3 reveals a symmetric distribution of door weights around its mean. If we can shift the process mean to the center of the specifications, it would bring a greater proportion of the door weights within the specifications. Thus, if our steel supplier turns down the thickness setting on his sheet rolling

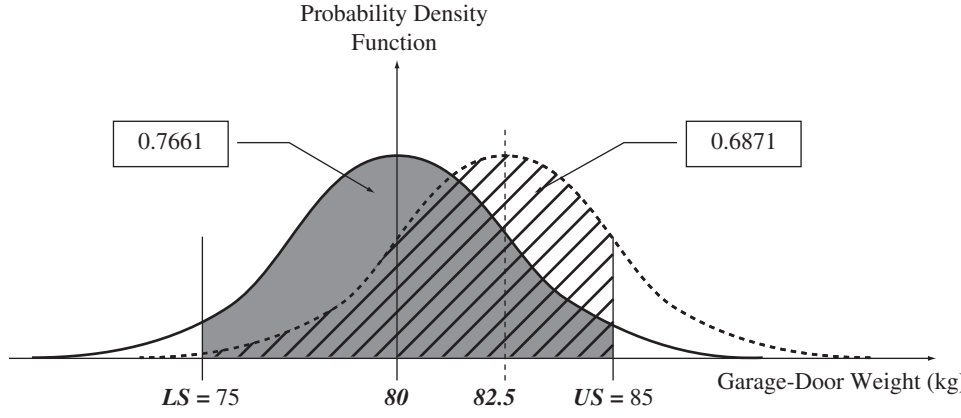


FIGURE 9.13 Process Improvement from the Mean Shift

mill, he may be able to reduce the average door weight down to $\mu = 80$ kg., thereby shifting the entire distribution of door weights to the left. When that happens, the reader may verify that the proportion of doors produced within the specifications increases to

$$\begin{aligned}\text{Prob}(75 \le \mathbf{X} \le 85) &= \text{Prob}(\mathbf{X} \le 85) - \text{Prob}(\mathbf{X} \le 75)\\ &= 0.88307 - 0.11693\\ &= 0.766141\end{aligned}$$

Figure 9.13 shows the improvement in the proportion meeting specifications by shifting the process mean from 82.5 to 80.

As we saw in Example 9.8, the process capability index C_{pk} increases from 0.1984 to 0.3968.

Thus, centering the process appropriately improves its capability. Any further improvement must come from reduction in the normal process variability (recall that the process is in control so there is no indication of abnormal variability).

9.5.2 Variability Reduction

Currently, there is too much variability in weights from one door to the next, as measured by the standard deviation estimated to be 4.2 kg. This lack of consistency may be due to any number of causes—perhaps the door fabrication equipment is too old, poorly maintained, and imprecise; perhaps the operator has not been trained properly; or perhaps the steel rolls delivered by the supplier are inconsistent from one batch to the next because of imprecision in the rolling mill.

If such causes of variability were corrected—through investment in better equipment, worker training, or supplier selection—the process output would be more consistent. In turn, that consistency would be reflected in a smaller standard deviation and a greater concentration of the frequency distribution closer to the mean. A greater fraction of output would fall within the specifications. In this case, note that reducing the process average is much easier and can be done quickly by the appropriate worker (or at least the supervisor). Reducing process variability, however, requires considerable time, effort, and investment in resources and is therefore management's responsibility. Sometimes, even reducing the process mean might require considerable effort. For example, reducing the average processing time in manufacturing or waiting time in a service operation usually requires considerable investment in process capacity.

EXAMPLE 9.11

Returning to our door weight problem, suppose we can also reduce the standard deviation of weights from its current estimate of 4.2 to 2.5 kg. Then we can verify that the proportion of the output meeting specifications will increase to

$$\text{Prob}(75 \le \text{X} \le 85) = 0.9544$$

with corresponding

$$C_p = (85 - 75)/[(6)(2.5)] = 0.67$$

Figure 9.14 shows improvement in the proportion of output meeting specifications that comes from reducing the process variability.

Suppose we would like 99% of our output to meet the specifications. How precise must our process be? From properties of normal distribution, we know that $z = 2.58$

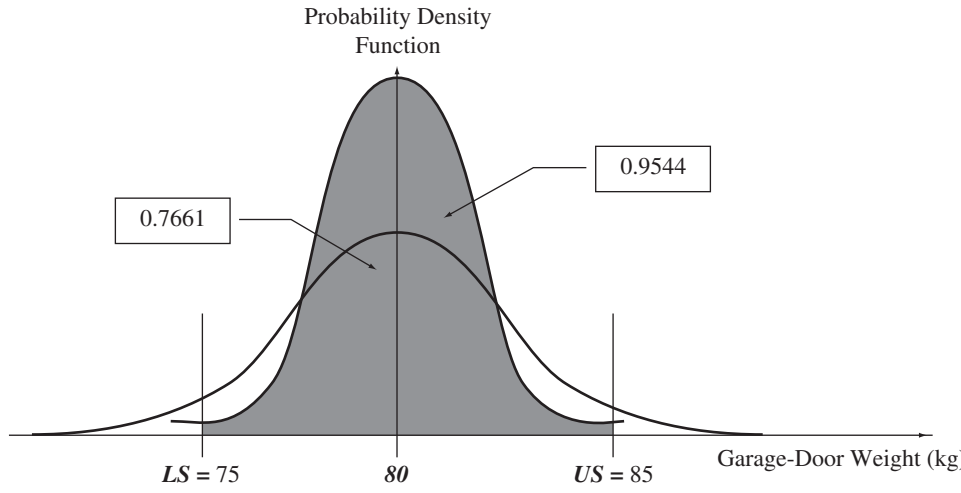


FIGURE 9.14 Process Improvement from Mean Shift and Variability Reduction

standard deviations on either side of the mean covers 99% of the area under the curve, so σ must be such that the upper and lower specifications are $z = 2.58$ standard deviations from the centered mean. In other words, we must have

$$2.58\sigma = 5$$

or

$$\sigma = 1.938 \text{ kg.}$$

which corresponds to

$$C_p = (85 - 75)/[(6)(1.938)] = 0.86$$

9.5.3 Effect of Process Improvement on Process Control

As the process capability is improved by shifting its mean μ or reducing its variability σ, process control limits must also be adjusted accordingly.

EXAMPLE 9.12

After only adjusting the process mean from 82.5 kg. down to 80 kg., the new control limits would be

$$80 \pm (3)(4.2)/\sqrt{5} = (74.36\,,\, 85.63)$$

Similarly, if we then also reduce the standard deviation from 4.2 to 2.5 kg., we need to revise control limits to

$$80 \pm (3)(2.5)/\sqrt{5} = (76.65\,,\, 83.35)$$

From then on, we would compare observed average weights of five doors against these new control limits to identify the presence of assignable causes of abnormal variability.

Thus, process control limits should be readjusted each time the process parameters are changed. Note that control limits on a more precise process are tighter since we expect

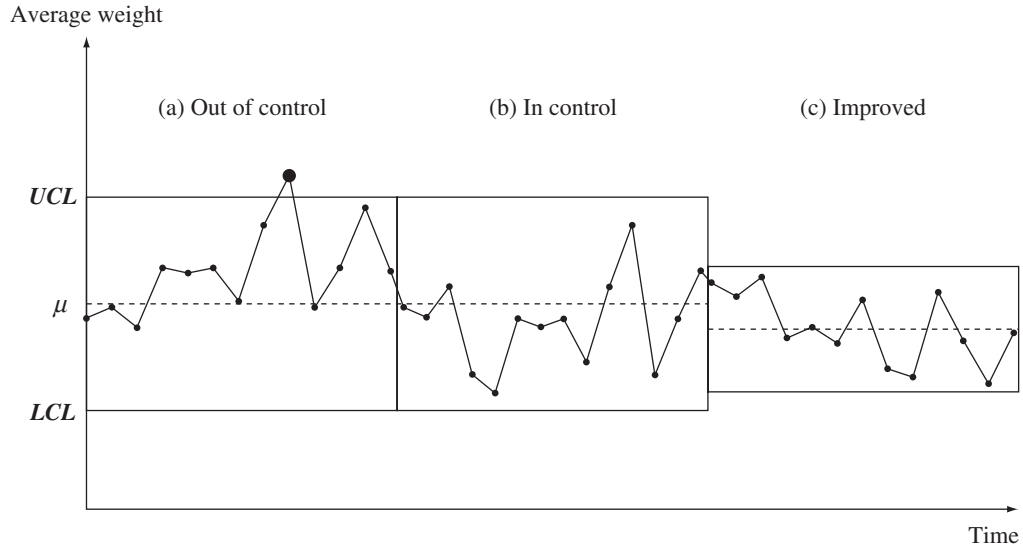


FIGURE 9.15 From Control to Capability Improvement

normal variability to be less. It is important to stress again that process control charts plot *sample averages*, whereas process capability refers to the ability of the process to meet specifications on *individual* units; the two should not be plotted on the same graph.

Figure 9.15 shows the progression in managing process variability from being (1) out of control to (2) in control by eliminating abnormal variability and, finally, to (3) greater capability through proper centering and reduced normal variability.

Although progression from (1) to (2) can be achieved in the short run, further improvement from (2) to (3) is a long-term journey. The next section indicates some steps in that direction through improved product and process design.

9.6 PRODUCT AND PROCESS DESIGN

Often, the sources of performance variability and poor quality can be traced to the poor design of the product and the process that produces it. In this section, we indicate a few general principles of design for minimizing the sources of process variability and its impact on product performance. The concept is simply that "an ounce of prevention is better than a pound of cure," to quote Benjamin Franklin.

9.6.1 Design for Producibility

We outline three general principles of product and process design aimed at minimizing chances of variability: product and process simplification, standardization, and mistake-proofing.

Simplification The objective here is to simplify product (or process) design so that it has fewer parts (or processing stages), which would then require fewer suppliers and reduced chances of confusion and error. If a product (or process) contains n parts (or processing stages) and each has probability p of performing successfully, then the probability that the *entire* product (process) will perform successfully—that *all* n parts will work—is p^n, which decreases geometrically as n increases, so that reducing n will improve its reliability.

Product simplification without foregoing product variety can be achieved through use of interchangeable parts and modular designs. For example, a sedan and minivan models of automobiles may share a common platform. Swatch watches often have identical internal parts and mechanism, while different dials and straps allow greater variety. They also simplify materials handling and inventory control. Recall from Chapter 7 the benefits of inventory pooling as a result of parts commonality, which reduces the amount of safety inventory needed.

Process simplification by eliminating non-value-adding steps in processing not only reduces the processing cost and flow time but also reduces opportunities for making mistakes. Minimizing the number of handoffs reduces chance of miscommunication that may lead to errors, for example in administering a wrong drug to a patient, or removing a wrong kidney in a surgery or implanting a lens in a wrong eye of a cataract patient. Complex accounting, medical, and military systems are increasingly vulnerable to failure. In general, "keep it simple, stupid!" (KISS) is an important design principle that requires ingenious and innovative ways of identifying opportunities to eliminate unnecessary, non-value-adding parts and processing steps.

Standardization Although product proliferation provides greater variety to customers, it increases the complexity of processing, which leads to higher cost, longer flow times, and lower quality. It is important to add only features that customers care about and are willing to pay for. Using standard, proven parts and procedures removes operator discretion, ambiguity, and chance of errors. Basic principles of using standard operating procedures are transferable across businesses. For example, hospitals can adopt communication strategies used by military personnel and airline pilots (e.g., flight check lists) to ensure patient safety and error-free healthcare. Surgeons may adopt time-outs just before surgery to confirm patient's name, date of birth and the procedure decided upon. Likewise, standard operating procedures simplify the tasks of recruiting and training employees and improve their performance consistency. As discussed in Chapter 2, flow shops producing limited variety products in high volumes enjoy low cost, short flow times, and consistent quality of output. Finally, even in service operations, as we saw in Chapter 8, reducing variability in processing times through standardization reduces customer waiting times, improving their perception of service quality.

Mistake-Proofing By minimizing the chance of human error, foolproofing a process improves product quality, reduces rework, and thus reduces both flow time and processing cost. Product design for ease of assembly is critical because assembly operations account for two-thirds of all manufacturing costs and are a major source of quality problems. Fasteners (e.g., screws and bolts), for instance, are widely known as potential problem sources and should be replaced with self-locking mechanisms. In product assembly, parts that have to be fitted together should be designed with either perfect symmetry or obvious asymmetry to prevent the possibility of misorientation. Workers and equipment should always have adequate clearance, unobstructed vision, and easy access to facilitate assembly. These principles of design for manufacturing (DFM) are well known in engineering literature and practice.

Techniques such as alphanumeric coding, color coding, and bar coding parts help make processing operations error resistant. Use of automation generally reduces labor costs, as well as chances of human error, and increases processing speed and consistency of the output. Electronic medical records have facilitated keeping patient information up-to-date as well as accurate, thereby minimizing the chance of wrong diagnosis and treatment. Bar coding has significantly reduced the billing errors at supermarket checkouts, as well as reducing the number of bags lost by airlines. Radio frequency identification (RFID) technology of electronic tagging is even more accurate,

fast, and robust (although more expensive) in tracking products through supply chains from manufacturing to point of sale, making them more transparent and reducing ordering errors.

Although many of these design principles to reduce variability and errors may appear obvious, their implementation in practice requires ingenuity, patience, and experimentation, with lot of communication, cooperation and contribution from workers, suppliers, and customers.

9.6.2 Robust Design

Until now, we focused on ways to eliminate assignable variability in the short run and reduce normal variability in the long run. Sometimes, however, variability reduction may not be possible or economical. An alternate approach to dealing with variability is through **robust design**. The idea is to *design the product so that its actual performance will not suffer despite any variability in the production process or in the customer's operating environment.* The goal is to develop a design that is robust in resisting effects of variability. For example, suppose we wish to design a cookie mix with the goal of optimizing the cookie quality in terms of flavor, texture, and taste. The recipe would involve specifying the type and amount of flour, sugar, and yeast in the mix, amount of water and butter to add, and the oven temperature and baking time. Variability in the cookie-making process arises from the package storage conditions and duration, customer errors in following the recipe, and the quality of the oven used. A robust design of the cookie mix would produce high quality cookies in spite of these sources of variability.

In general, product performance is determined by internal (process-related) and external (environment-related) noise factors along with its own design parameters. The designer's goal is to identify a combination of design parameters that will protect product performance from the internal and external noise factors which it may be subjected to. In statistically planned experiments, different combinations of design factors are tested in conjunction with combinations of different levels of noise factors. The challenge is to identify the right combination of design parameters (without trying them all) that works well, on average, in spite of noise factors. More details may be found in Taguchi and Clausing (1990).

9.6.3 Integrated Design

The goal of product or process design is to develop high-quality, low-cost products, fast. Design is a critical part of the product life cycle, because typically 70 to 90% of product cost is locked in at the design stage, while 60 to 80% of product problems can be traced to poor design. Conventional design process is sequential, like a relay race: Marketing determines what customers want, then designers design the product they think will meet customer requirements, and "throw" the design "over the wall" to manufacturing. Each handoff invites the possibility of miscommunication, or a gap mentioned in Section 9.1: Design may not satisfy customer needs on the one hand, and be producible on the other. Any problem at the manufacturing stage leads to rework and redesign by "going back to the drawing board." This iterative process increases product development cost, delays new product introduction, and may result in loss of first-mover advantage and competitive position.

In contrast, **integrated design** process works in parallel like a football team or an orchestra that involves customers, designers, suppliers, and producers early on to jointly conceive, develop, and implement product design and development programs to meet customer requirements. This requires breaking down barriers across functional silos, encouraging cross-functional training, close communication, and teamwork.

Early involvement by everyone concerned results in fewer revisions, shorter development time, and lower cost. Better communication among players minimizes misinterpretation, facilitates understanding one another's problems, limits and priorities, and results in smoother transfer of designs between stages. Quality function deployment (QFD) mentioned in Section 9.1 is an example of integrated design.

Although integrated design (also known as parallel or concurrent design) makes perfectly good sense in principle, its implementation is challenging in practice. There are several organizational issues involved: Different players speak different languages, there may be turf wars between groups, coordination among diverse groups is often complicated, incentives have to be team- rather than individual-based. And above all, any change from the conventional way is always difficult. Introduction of the Boeing 787 Dreamliner illustrates difficulties involved in design and development of new products. While revolutionary in terms of fuel efficiency, speed, range, and passenger comfort (lighting, noise, humidity, and air quality), coordinating an international team of designers and suppliers proved challenging, which resulted in repeated delays in design, production, and delivery of the aircraft.

We conclude the chapter by listing some general principles of total quality management (TQM), which emphasizes the holistic nature of quality aimed at product and process design and control to minimize gaps and variability, ultimately leading to customer satisfaction.

- Customer focus: Identify customer needs
- Integrated design: Involve the entire organization
- Build-in quality: Emphasize early prevention
- Supplier involvement: Ensure cooperation, commitment, and trust
- Employee involvement: Empower employees for local process control
- Continuous improvement: Manage by facts, data, and measurement
- Long-term perspective: Invest in resources, equipment, and training

Summary

In this chapter, we emphasized how performance variability over time—and not just its average—is an important determinant of customer satisfaction. We first presented some simple graphical and statistical tools—such as check sheets, Pareto charts, and histograms—for documenting, organizing, and summarizing information about observed variability. We then extended this *static* analysis to *dynamic* tools such as run charts, multi-vari charts, and, most importantly, control charts to track performance variability over time.

We outlined the feedback control principle for monitoring and acting on the observed variability over time. We learned that a deviation in process performance from its expected value may be due to normal or abnormal variability. We studied process control charts as a prime application of this principle that enables us to detect the presence of abnormal variability. Setting up a control chart involves (1) estimating the average performance and normal variability around it and (2) establishing limits of acceptable variability in performance around the average. Implementing a control chart involves (1) monitoring and plotting the actual performance against these limits and (2) signaling the presence of abnormal variability that warrants an action when these limits are exceeded. We indicated cause–effect diagrams and scatter plots as simple tools for identifying and correcting causes of abnormal variability. Thus, the goal of process control is to detect when a process goes "out of control" and eliminate causes of abnormal variability to bring the process back "in control" so it displays only normal variability, which signifies a state of internal stability.

We then studied process capability in terms of its ability to meet external customer requirements. We

computed the fraction of the output that meets customer specifications, the process capability ratio, and sigma capability as three related measures of process capability. All of them try to quantify the magnitude of process variability in relation to customer specifications. We outlined few strategies for improving process capability by reducing its normal variability through better product and process design to facilitate error-proof processing through simplification, standardization, and mistake-proofing. Finally, we indicated the concept of robust design of products, which desensitizes their performance to sources of process variability.

Although performance variability will always plague every process, it becomes troublesome when it leads to process instability, lower capability, and customer dissatisfaction. In this chapter, our goal has been to study how to measure, analyze, and minimize sources of this variability so as to improve consistency in product and process performance, ultimately leading to total customer satisfaction and superior competitive position.

Key Equations and Symbols

(Equation 9.1) $LCL = \mu - z\sigma$ and $UCL = \mu + z\sigma$
(generic control limits)

(Equation 9.2) $LCL = \overline{\overline{X}} - zs/\sqrt{n}$ and $UCL = \overline{\overline{X}} + zs/\sqrt{n}$
(average or $\overline{X}$ control chart limits)

(Equation 9.3) $LCL = \overline{R} - zs_R$ and $UCL = \overline{R} + zs_R$
(range or R control chart limits)

(Equation 9.4) $LCL = \overline{p} - z\sqrt{\overline{p}(1 - \overline{p})/n}$ and $UCL = \overline{p} + z\sqrt{\overline{p}(1 - \overline{p})/n}$
(fraction defective or p control chart)

(Equation 9.5) $LCL = \overline{c} - z\sqrt{\overline{c}}$ and $UCL = \overline{c} + z\sqrt{\overline{c}}$
(number of defects or c control chart limits)

(Equation 9.6) $C_{pk} = \min[(US - \mu)/3\sigma, (\mu - LS)/3\sigma]$
(process capability ratio)

(Equation 9.7) $C_p = (US - LS)/6\sigma$
(process capability ratio for a centered process)

(Equation 9.8) $S = \min[(US - \mu)/\sigma, (\mu - LS)/\sigma]$
(sigma capability)

(Equation 9.9) $S = (US - LS)/2\sigma$
(sigma capability of a centered process)

where
LCL = Lower control limit
UCL = Upper control limit
μ = Process mean
σ = Process standard deviation
z = Measure of tightness of control
$\overline{X}$ = Sample average
$\overline{\overline{X}}$ = Grand average of sample averages
s = Sample standard deviation
R = Sample range
$\overline{R}$ = Average range
s_R = Standard deviation of ranges
$\overline{p}$ = Proportion defective
$\overline{c}$ = Average number of defects per flow unit
C_{pk} = Process capability ratio (for noncentered process)
C_p = Process capability ratio (for centered process)
US = Upper specification
LS = Lower specification

Key Terms

- 80-20 Pareto principle
- Abnormal variability
- Cause–effect diagram
- Check sheet
- Control band
- Control chart
- Control limits
- Feedback control principle
- Fishbone diagram
- Fraction defective (or p) chart
- Histogram
- Integrated design
- Ishikawa diagram
- Lower control limit (*LCL*)
- Lower specification (*LS)*
- Multi-vari chart
- Normal variability
- Number of defects (or c) Chart
- Pareto chart
- Plan-Do-Check-Act (PDCA) cycle
- Process capability
- Quality function deployment (QFD)
- Quality of conformance
- Quality of design
- *R*-Chart
- Robust design
- Run chart
- Scatter plot
- sigma [capability] measure
- Six-sigma process
- Total Quality Management (TQM)
- Type I (or α) error
- Type II (or β) error
- Upper control limit (*UCL)*
- Upper specification (*US*)
- *X*-Bar Chart

Discussion Questions

9.1 Discuss with three examples from everyday life where variability in product or process performance leads to customer dissatisfaction even though the average performance is considered good.

9.2 Suppose you are managing a grocery store and would like to provide a first-rate shopping experience to your customers. Outline how you would go about determining factors that are important to them, how well you are doing in meeting their needs and expectations, and how you can improve your operations to give them total customer satisfaction. Specifically, discuss which of the tools that you learned in this chapter can be used and how.

9.3 In operating an airline, on-time performance is a critical measure that customers value. Suppose you plot a histogram of actual arrival and departure times in relation to the scheduled times. What information will it provide that you can use to improve the process?

9.4 What information do run charts, multi-vari charts, and control charts provide in addition to that contained in a histogram that shows variability in process performance?

9.5 Give three everyday life examples of situations where the feedback control principle can be applied for collecting information and making decisions.

9.6 What are two main types of process variability? How can we identify and remove the sources of this variability?

9.7 What factors should one consider in setting control limits? What are the trade-offs involved in determining the width of a control band?

9.8 How can the process be "in control" but have dissatisfied customers? It sounds like "the operation was successful, but the patient died." Comment on this apparent paradox.

9.9 What are two concrete ways of measuring process capability? How are they related? How can they be improved?

9.10 What does six-sigma capability mean? Why is it important to insist on such high standards?

9.11 It has been observed that in the airline industry, baggage handling is about a four-sigma process, whereas frequency of airline fatalities corresponds to a seven-sigma capability. How can two processes within the same industry be so different?

9.12 Give three examples of improving process capability through better design.

Exercises

***9.1** Costello Labs supplies 500-cc bottles of treated Elixir plasma solution to Mercy Hospital. Several factors are important in assessing plasma quality, such as purity, absence of AIDS or hepatitis virus, and bacterial count. The most important quality characteristic, however, is protein concentration. Protein concentration is measured by a sophisticated electronic process known as electrophoresis. American Medical Association (AMA) standards specify that a 500-cc plasma bottle should contain between 30 and 35 grams of protein. Both concentrations under and over this range may be hazardous to a patient's health.

Hospital administrators have instructed Costello Labs to straighten out its plasma production operation and to demonstrate evidence of tighter process controls prior to the renewal of its supply contract. Costello's plasma production equipment consists of a protein injector and a mixer that determine protein concentration in each bottle. Process capability depends on the precision of these pieces of equipment.

a. Suppose that the hospital and the lab have agreed that at least 98% of the plasma bottles supplied by Costello should conform to AMA specifications (i.e., should contain between 30 and 35 grams of protein). Determine the following:

- The precision of Costello's protein injector as measured by σ
- The standard deviation of the amount of protein that it must inject into each bottle in order to produce 98% of process output within the specifications

Also compute the corresponding process capability ratio C_p.

b. Costello Labs production manager Phil Abbott wants to establish statistical process control charts to monitor the plasma-injection process. Set up these control charts based on average protein readings taken from randomly selected samples of 12 bottles from each batch.

9.2 Natural Foods sells Takeoff, a breakfast cereal, in one-pound boxes. According to Food and Drug Administration (FDA) regulations, a one-pound box must contain at least 15.5 ounces of cereal. However, Natural Food's box-filling process is not perfect: its precision, expressed in terms of the standard deviation of the weight of a one-pound box filled, is 0.5 ounces.

a. Where should Natural Foods center its process in order to ensure that 98% of boxes filled meet FDA requirements? What proportion of boxes would be overfilled beyond 16 ounces?

b. While underweight boxes might prompt FDA action, overweight boxes certainly cost Natural Foods in terms of higher material costs. Therefore, quality control manager Morris Nerdstat wants to monitor the cereal-filling process in order to ensure that its mean does not deviate from the level established in Part a. He plans to weigh nine randomly selected boxes at regular time intervals and plot the average weight on a chart. At one point, he finds an average weight of 15.9 ounces. The company's legal staff is pleased that this performance is better than the FDA requirement of 15.5 ounces. What action, if any, should Nerdstat take?

c. What targets (in terms of the mean and the standard deviation) would result in the process with six-sigma capability?

***9.3** In measuring and evaluating the quality of banking services, analysts have found that customers regard accuracy, timeliness, and responsiveness as the most important characteristics. Accordingly, First Chicago Bank constantly monitors and charts almost 500 performance measures of these quality characteristics. Accuracy, for example, is measured by the error/reject rate in processing transactions, timeliness by delays in collecting funds, and responsiveness by speed in resolving customer inquiries or complaints. For each measure, First Chicago also sets a level called Minimal Acceptable Performance (MAP), which serves as an early warning signal to management, as a basis for comparison with the banking-industry competition, and as a constantly upgraded goal for ensuring continuous improvement of service quality.

Over one six-month period, First Chicago recorded, on a weekly basis, errors per thousand items processed in all types of collection transactions. The resulting 26 numbers were as follows: 0, 2, 0, 17, 2, 4, 0, 2, 1, 0, 0, 5, 6, 5, 15, 5, 10, 5, 2, 2, 0, 2, 0, 0, 0, and 1. The Bank Administration Institute reports that the average error rate for such transactions is 1.5%.

a. Determine the appropriate process control limits on the fraction of transactions in error, or the number of errors per thousand transactions.

b. By way of comparison, plot the observations, process average, control limits, and industry standard. Is First Chicago's process in control? How does its performance compare with the industry standard?

9.4 Government regulations mandate that Belgian chocolate bars sold in packages of ½ kilogram cannot weigh less than ½ kilogram, the specified weight on the package. If regulations are violated, the company is fined. The chocolate machine at the Cote d'Or chocolate company fills packages with a standard deviation of 5 grams regardless of the mean setting. To be sure that government regulations are met, the operator decides to set the mean at 515 grams.

a. To check if the process is in control, the operator plans to take samples where each sample consists of 25 chocolate bars. The average weight of this sample of 25 bars is used to determine if the process is in control. Following industry practice, what control limits should the operator use?

b. If the process is in control, approximately what fraction of chocolate bars will weigh less than 500 grams (this is the fraction that would violate government regulation)?

c. Clearly, producing an excess average chocolate weight of 15 grams just in order to prevent regulation fines is costly in terms of chocolate "given away for free." Cote d'Or management wants to reduce the average excess weight to 3 grams while staying in line with regulations "practically always," which means 99.87% of the time. In what sense will this require improved process technology? Give an explanation in words as well as a specific numeric answer.

***9.5** Consider a machine producing drive shafts with a mean diameter of 6 cm and a standard deviation of 0.01 centimeters. To see if the process is in control, we take samples of size 10 and evaluate the average diameter over the entire sample. Customer specifications require drive shafts to be between 5.98 and 6.02 centimeters. Establish the appropriate control limits.

9.6 If product specifications remain unchanged, as the sigma capability of a process improves from being a 1-sigma to a 4-sigma process, what is the effect on the range between the control limits (UCL – LCL)? Does it become narrower, wider, or remain unchanged? Why?

9.7 A bank has recently taken over the billing function of a company. An agreement stipulates that the bank should process 99.2% of the bills within 4 hours. A study of the current process indicates that the processing time averages 2.2 hours per bill with a standard deviation of 0.8. A process improvement team has suggested a change. Experiments indicate that the change will lower the average processing time to 2 hours but raise the standard deviation to 1.2. Should this change be implemented? Why?

***9.8** Balding Inc. is the official producer of basketballs used in NBA tournaments. The Dallas Mavericks have placed a large order for Balding's Fusion basketballs of 29.5-inch diameter, which feature exclusive micropump technology. Jennifer Boling, the Mavericks' head of basketball operations, says she will accept only basketballs within 0.2-inch tolerance (i.e., those with diameters between 29.3 and 29.7 inches). Balding's production manager sets the mean of its basketball manufacturing process at 29.5 inches and would like to ensure that 95% of its output will meet Boling's requirements. What should be Balding's process capability ratio?

9.9 Evaluate with explanation the following statements, circle the appropriate response, and explain briefly.
- **a.** Suppose the control limits for a process are set at 3 standard deviations from the mean. If the process is in control, 99.73% of its output will meet the specifications. True or false?
- **b.** As the sample size increases, the upper control limit for a process should be decreased. True or false?
- **c.** Suppose control limits for a process are set at 3 standard deviations from the mean. If the process is in control, the probability of observing a sample outside the control limits is independent of the sigma capability of the process. True or false?
- **d.** A six-sigma process is always "in control." True or false?

9.10 Specify appropriate control chart to monitor performance in the following instances in assessing performance of an airline.
- **a.** Flight delay from the announced arrival time
- **b.** Fraction of flights delayed out of twenty flown
- **c.** Number of bags lost per 1000 flown
- **d.** Number of customer complaints received per month
- **e.** Length of time required to resolve a complaint
- **f.** Number of flights delayed per month due to mechanical breakdowns

Selected Bibliography

Crosby, P. B. *Quality is Free.* New York: McGraw-Hill, 1979.

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

Feigenbaum, A. V. *Total Quality Control.* New York: McGraw-Hill, 1961.

Garvin, D. A. *Managing Quality.* New York: Free Press, 1988.

Grant, E. L., and R. S. Leavenworth. *Statistical Quality Control.* 6th ed. New York: McGraw-Hill, 1988.

Hauser, J., and D. Clausing. "The House of Quality." *Harvard Business Review* 66, no. 3 (May–June 1988): 63–73.

Joiner, B., and M. Gaudard. "Variation, Management, and W. Edwards Deming." *Quality Progress*. Special Issue on Variation, December, 1990.

Juran, J. M., and F. M. Gryna. *Quality Control Handbook.* 4th ed. New York: McGraw-Hill, 1988.

Juran, J. M., and F. M. Gryna. *Quality Planning and Analysis.* 2nd ed. New York: McGraw-Hill, 1980.

Ott, E. R., and E. J Schilling. *Process Quality Control: Troubleshooting and Interpretation of Data.* 2nd ed. New York: McGraw-Hill, 1990.

Taguchi, G., and D. Clausing. "Robust Quality." *Harvard Business Review* 68, no. 1 (January–February 1990): 65–75.

Wadsworth, H. M., K. Stephens, and B. Godfrey. *Modern Methods for Quality Control and Improvement.* New York: John Wiley & Sons, 1986.

Wheeler, D. J. *Understanding Variation: The Key to Managing Chaos*, Knoxville, TN: SPC Press, 2000.

PART IV

Process Integration

Chapter

10

Lean Operations: Process Synchronization and Improvement

INTRODUCTION

Following the early lead of Toyota, many industries have used the principles of lean operations to improve performance in terms of cost, quality, and response time within their plants and supply chains. During the 1980s, the Toyota Production System (TPS) garnered increasing attention to understand the growing success behind Toyota's and other Japanese manufacturing industries in global markets. To generalize the principles behind TPS to other manufacturing industries, the term *lean* was coined. Lean was chosen to highlight the principles of limiting inventory, excess workers, or "waste," as opposed to other auto manufacturers' "buffered" approaches (Hopp and Spearman, 2004). Doig et al. (2003) discuss how some airlines have used lean operations techniques to improve aircraft and component turnaround times during maintenance by 30 to 50 percent and maintenance productivity by 25 to 50 percent. Being lean has also yielded dramatic improvements for the venerable French luxury-goods house Louis Vuitton. In 2005, it took 20 to 30 craftsmen to put together each "Reade" tote bag over the course of about 8 days. After a lean transformation in 2006, "clusters of six to 12 workers, each of them performing several tasks, can assemble the $680 shiny, LV-logo bags in a single day" (Passariello 2006).

Over the last decade, lean thinking also has diffused from manufacturing to service operations. In 2004, the Indian software services provider Wipro Technologies quietly launched a pilot "lean" initiative: an endeavor that attempted to translate ideas on lean production from manufacturing to software services (Staats and Upton, 2009). By June 2007, Wipro had 772 lean projects completed or underway. Staats and Upton show that lean projects performed better, and with lower variation than a matched comparison set in many, but

not all cases. This assessment suggests the applicability of manufacturing-based principles to a fast-moving, high-tech service industry.

Similar improvements have also been observed in other service industries. According to Swank (2003), Jefferson Pilot Financial, a full-service life insurance company, applied lean production techniques to the process of handling applications from its premier partners. The result was a 26 percent reduction in labor costs and a 40 percent reduction in the rate of reissues due to errors. These outcomes increased new annualized premiums collected in the company's life insurance business by 60 percent over two years. According to Wysocki (2004), Allegheny General Hospital in Pittsburgh used the quality at source idea from lean operations to cut infections by 90 percent within 90 days.

In this chapter, we examine how lean operations can improve performance in terms of cost, quality, and response time in organizations. What are the basic principles of lean operations? How can they be used to improve flows within a single site? How can these ideas be applied to improve flows across the supply chain?

In Section 10.1, we discuss how single sites (e.g., factories) and supply chains can be represented as processing networks. In Section 10.2, we characterize the ideal performance of a processing network in meeting customer requirements in terms of flow synchronization and cost efficiency. In Section 10.3, we view any deviation from this ideal as waste and examine its sources and consequences. We define the goal of process improvement as bringing the process performance closer to the ideal by identifying and eliminating waste. In Section 10.4, we study methods of lean operations designed to improve plant-level performance by increasing flexibility, reducing variability, and improving information and material flows. In Section 10.5, we extend lean ideas toward the goal of achieving synchronization and efficiency across the entire supply chain. Finally, in Section 10.6, we look at the process of improvement and compare two general approaches to attaining ideal performance: continuous improvement and process reengineering. We also indicate the role of benchmarking in setting process-improvement goals and the importance of managing the organizational change that always accompanies process improvement.

10.1 PROCESSING NETWORKS

As discussed in Chapter 1, any organization can be viewed as a business process that transforms inputs into outputs to satisfy customer needs. A firm satisfies customers by providing them what they want, when they want it, and where they want it at a price they are willing to pay. Theoretically, satisfying all these criteria would mean developing, producing, and delivering individually customized products of the highest quality, in the shortest time, and at the lowest cost. In reality, given the firm's capabilities and constraints, trade-offs must be considered. In most industries, as discussed in Chapter 2, there exists an operations frontier in the competitive product space defined by the four product attributes—cost, quality, variety, and response time. This frontier reflects the optimal trade-offs given the current state of technology and management practices. Competition forces firms operating below the industry's operations frontier to improve and move toward the frontier. World-class firms already operating at the frontier can stay ahead of competitors only by improving and pushing the frontier further. Thus, firms at every level have the scope and necessity to improve process performance along the four dimensions that customers value.

This chapter extends the concepts and principles developed thus far for individual processes to improve performance of a **processing network** that *consists of information and material flows of multiple products through a sequence of interconnected processes.* Figure 10.1 illustrates two product flows through a typical processing network. The overall goal of this network is to satisfy customer demand in the most economical way

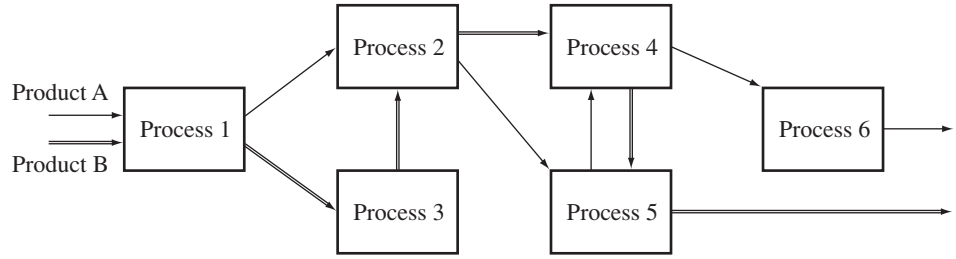


FIGURE 10.1 Product Flows in a Processing Network

by producing and delivering the right products, in the right quantities, at the right times, to the right places. It requires the synchronization of flows between processes in a cost-effective manner.

Plants and Supply Chains Our discussion focuses on performance at two different levels—plant and supply chain. A **plant** is *any singly owned, independently managed and operated facility, such as a manufacturing site, a service unit, or a storage warehouse.* A **supply chain** is *a network of interconnected facilities of diverse ownership with flows of information and materials between them*. It can include raw materials suppliers, finished-goods producers, wholesalers, distributors, and retailers. For example, the supply chain that makes a detergent available in a supermarket includes chemical plants, warehouses for storing chemicals, factories for producing and packaging detergents, distributors, wholesalers, and, finally, retailers. Each facility represents a plant, whereas all of them together form the detergent supply chain.

If we view the facilities in greater detail, each plant in the supply chain is also a processing network. Within the detergent maker's factory, the purchasing, production, storage, and shipping departments are all processes, each handling a variety of detergents and cleaners. In this chapter, we first examine how to manage processing network operations within a given plant and then extend the key principles to coordinate the operations of the entire supply chain. The core ideas that apply to both levels of operation are the same and draw on the process-improvement levers that we have discussed in earlier chapters. However, the operational details differ because of differences in scale, scope, geographical dispersion, and incentives of the diverse process owners involved.

10.2 THE PROCESS IDEAL: SYNCHRONIZATION AND EFFICIENCY

Customers want a wide variety of high-quality products from convenient locations at low prices. Performance of an **ideal process**—*a process that achieves synchronization at the lowest possible cost*—can thus be summarized in terms of two closely related operating characteristics:

1. **Process synchronization** refers to *the ability of the process to meet customer demand in terms of their quantity, time, quality, and location requirements.*
2. **Process efficiency** is *measured in terms of the total processing cost.*

The Four "Just Rights" of Synchronization A well-synchronized detergent supply chain produces and delivers the right quantities of defect-free boxes of the detergent to widely dispersed supermarkets so that just enough is available to satisfy all customer demand without delay. For manufactured goods, customer demand can always be satisfied by producing in advance and carrying large inventories of all products, of verified quality, in all locations. This approach, however, is not synchronized with demand

and not very efficient. We therefore define a perfectly synchronized process as one that is lean in that it develops, produces, and delivers the following only on demand:

- Exactly *what* is needed (not wrong or defective products)
- Exactly *how much* is needed (neither more nor less)
- Exactly *when* it is needed (not before or after)
- Exactly *where* it is needed (not somewhere else)

A perfectly synchronized process always supplies just the right *quantity* of the right *quality* product, at just the right *time*, and in just the right *place*—just as desired by customers. These four "just rights" of synchronization form the core of the **just-in-time (JIT)** paradigm. Just-in-time refers to *an action taken only when it becomes necessary. In manufacturing, it means production of only necessary flow units in necessary quantities at necessary times.*

These four criteria define the ultimate in process quality, flexibility, capacity, and speed. Producing any product without defects requires the process to be extremely versatile and precise. The ability to produce any desired quantity requires flexibility to produce one unit at a time. In order to satisfy demand arising at any time—without entailing inventories—a process must have instant, complete, and accurate information on demand and must be able to react by producing and delivering instantly as well. An ideal process can satisfy all these requirements and do so at the lowest possible cost. In short, an ideal process is infinitely capable, flexible, fast, and frugal.

Synchronized Networks This concept of an ideal process extends naturally to a *network* of processes—once we recognize that in such a network, the outflow of one process (a supplier) is the inflow to another (a customer). Perfect synchronization of an entire network of processes requires precise matching of supply and demand of various flow units at each processing stage. It means that each stage must satisfy—precisely—the quality, quantity, time, and place requirements of the next stage.

We can define synchronization at the level of an individual process (as a network of activities), a plant (as a network of processes), or a supply chain (as a network of plants). In each case, the goal of ideal performance requires that individual processing stages be capable, flexible, fast, and frugal. Synchronization requires all stages to be tightly linked in terms of the flow of information and product. The result is a precisely balanced system of inflows and outflows at all stages through which units flow smoothly and continuously without disruption or accumulation along the way. In particular, for a perfectly synchronized process, the output of every stage will precisely match (external) end-customer demand. In an ideal network, this synchronization of processing stages is achieved at the lowest possible cost.

Although the ideal may seem unattainable in a practical sense, the long-run goal—and challenge—of process management should be to approach this ideal by improving products, processes, and practices. In the next section, we examine the causes and consequences of failure to attain the ideal.

10.3 WASTE AND ITS SOURCES

It is important to focus on the ideal because anything short of ideal performance represents an opportunity for us to improve the process—or for the competition to move in. Operationally, low efficiency is reflected in high processing costs. Lack of synchronization manifests itself in defective products, high inventories, long delays, or frequent stockouts.

Sources of Waste Regarding any deviation from the ideal as **waste**, we paraphrase the goal of process improvement as the elimination of all waste. Thus, waste means

producing inefficiently, producing wrong or defective products, producing in quantities too large or too small, and delivering products too early or too late—that is, failing to match customer demand most economically. Taiichi Ohno, the main architect of the Toyota Production System, classified seven types of waste in manufacturing (1988):

- Producing defective products
- Producing too much product
- Carrying inventory
- Waiting due to unbalanced workloads
- Unnecessary processing
- Unnecessary worker movement
- Transporting materials

All this waste results in high costs, low quality, and long response times, ultimately leading to customer dissatisfaction and loss of business to the competition. Producing defective units results not only in unhappy customers but also in additional cost and time required to receive, inspect, test, rework, and return those units. Producing too much or too early builds up excess inventory, which increases holding costs (including the costs of capital, storage, and possible obsolescence) and the unit flow time. In turn, long flow times mean delays in responding to changes in customer tastes and in getting new product designs to market. In processing networks, inventory buffers between stages increase total flow time, thus delaying feedback on quality problems, obstructing traceability of root causes, and diffusing accountability for errors.

Producing too little or too late results in stockouts, delays, and increased expediting costs. In processing networks, insufficient inflows starve some stages, resulting in idleness and inefficient utilization of resources. Finally, delivering wrong products to wrong places creates excess inventories of wrong products, shortages of right ones, or both. Corrective transfers then result in additional costs and delays.

The sources of all this waste can ultimately be traced to underlying process imperfections, demand and supply variability, or management practices discussed throughout this book. We saw in Chapter 4 that non-value-adding activities, such as transportation, movement, inspection, and rework, increase theoretical flow time and processing costs. Similarly, we learned in Chapters 5 and 8 that insufficient capacity at bottlenecks reduces process throughput and increases waiting time. In Chapter 6, we observed that a lack of flexibility to switch between products, measured in terms of fixed setup (or changeover) costs, necessitates producing in large batches even though demand is continuous, a mismatch giving rise to cycle inventories. Likewise, we found in Chapter 7 that stochastic variability in supply and demand, together with long and uncertain lead times, requires us to hold safety inventory to protect against stockouts. If demand is predictable and all stages in a supply chain are both flexible in processing different products and predictable in terms of operating without variability, buffer inventories are unnecessary and flows synchronized. As discussed in Chapter 8, it is the variability in demand and processing times that causes both waiting and inflow inventory, thereby requiring safety capacity at an added cost. In Chapter 9, we saw that insufficient process capability in terms of high normal variability results in defective units. Meanwhile, abnormal variability in terms of process instability over time necessitates expensive process control. Finally, lack of synchronization from delivering wrong products to wrong locations is often due to inadequate transmission of information and materials through the network.

Waste Elimination Cycle and safety inventories, safety capacity, and non-value-adding activities including transportation, inspection, rework, and process control are short-term tactical actions that process managers take in order to work with imperfect

processes suffering from inflexibility, variability, and inefficient logistics. In the short term, we accept these process limitations as given and try to deal with them by processing in batches, incorporating safety stocks and safety capacity, monitoring processes, and correcting defects. All these measures, however, increase total cost, inventory, and flow time, resulting in less-than-ideal performance.

A long-term strategy is to improve the underlying process to make it more flexible, predictable, and stable, which eliminates the need for such temporary measures as batch processing, safety allowances, and process control. For example, recall from Chapter 6 that the optimal batch size (hence the cycle inventory and flow time) are directly related to the (square root of) setup cost. A long-term solution to reducing cycle inventories is to reduce the setup cost itself (i.e., improve process flexibility) so that it is economical to produce various products in small batches. Such an action improves synchronization by reducing cycle inventory and flow time.

Similarly, recall from Chapters 7 and 8 that the amount of safety inventory and safety capacity needed to provide a given level of service depends directly on the degree of variability in the system. A long-term solution calls for making flows more regular and predictable—in other words, for reducing variability. In doing so, we reduce the required safety cushion—thereby reducing cost while improving synchronization.

In Chapter 9, we identified two principles governing the relationship between variability and process stability. In the short run, the following holds:

1. Greater normal variability (i.e., lower process capability) results in more defective products
2. More abnormal variability (i.e., greater process instability) requires tighter process control to maintain stability

It follows, then, that reducing variability (by increasing process capability and stability) does the following:

1. Decreases the number of defective products, inspection, and rework
2. Reduces the need for online process control, thereby improving synchronization and reducing overall cost

Thus, the long-run goal of process improvement is to identify and eliminate the root causes of waste rather than to compensate for them with short-term solutions. The idea is to diagnose and remove the roots of an ailment, seeking a permanent cure, rather than superficially treating symptoms with temporary fixes.

The River Analogy Figure 10.2 illustrates the concept of waste and its sources by using a river analogy that has been popularized in the literature on the Toyota Production System. Visualize process imperfections—defective materials, machine breakdowns, long setup times, unreliable suppliers, inefficient layouts—as rocks lying on a riverbed. The water level in the river represents waste in the form of short-term measures such as excess cycle and safety inventories, safety capacity, time allowance, safety quality, inspection, rework, and process control. They provide an operating cushion to facilitate smooth sailing for the process manager despite underlying problems. The appropriate long-term response is to uncover and remove these rocks so that we can sail smoothly even in shallow water (which symbolizes lean operations). Three factors, however, impede us from achieving the long-term solution: (1) a high water level covers up rocks, reduces problem visibility, and clouds the root causes on the bottom; (2) smooth sailing because of the safety cushion dampens our incentives to look for root causes; and (3) lack of problem-solving skills makes it difficult to eliminate the root causes. The challenge of process management is to overcome these three obstructions and bring actual performance closer to the ideal. The river analogy suggests lowering

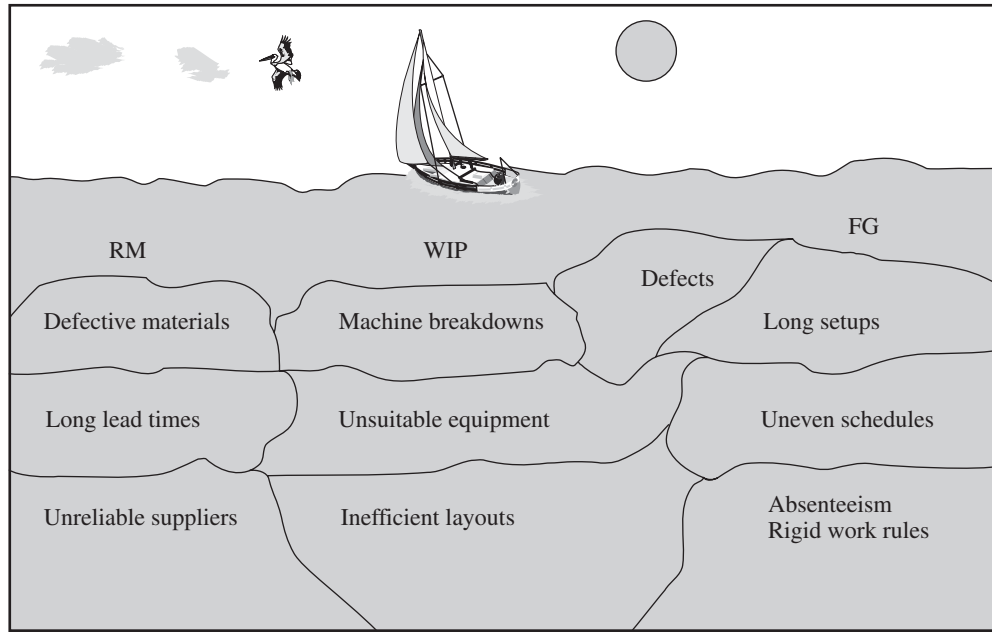


FIGURE 10.2 The River Analogy: Waste and Its Sources

the water level slowly until the top rocks are visible. Eliminating these rocks now provides smooth sailing with a lower level of water. The pressure for improvement is maintained by lowering the water level further until more rocks become visible. As rocks are constantly eliminated, a low level of water is sufficient to provide smooth sailing.

In the next two sections, we examine specific methods for improving process synchronization and efficiency, first within a plant and then within an entire supply chain. Although the operational details in the two contexts are different, the basic principles are the same. To improve process synchronization, we need to do the following:

- Synchronize flows of material and information
- Increase resource flexibility
- Reduce process variability

To improve process efficiency, we need to do the following:

- Reduce processing cost and flow time

These improvements at the plant and supply chain level require a long-term investment in the process, including equipment, technology, workers, and suppliers.

10.4 IMPROVING FLOWS IN A PLANT: BASIC PRINCIPLES OF LEAN OPERATIONS

Any plant, whether a manufacturing or a service facility, is a network of processing stages through which materials or customers flow before emerging as finished products or serviced customers. An ideal plant is synchronized and efficient: the outflow of each stage meets—precisely and economically—the inflow requirements of the next, without defects, inventories, delays, or stockouts. Methods for achieving efficiency and synchronization within a plant have been discussed in the operations management literature under such headings as lean operations, just-in-time production, zero inventory program, synchronous manufacturing, agile manufacturing, and the Toyota Production System (TPS).

According to its main architect, the basic objective of TPS is "to shorten the time it takes to convert customer orders into deliveries" (Ohno, 1988). TPS is, in a sense, like Mr. Ohno says:

> "making a factory operate for the company just like the human body operates for the individual. The autonomic nervous system responds even when we are asleep. The human body functions in good health when it is properly cared for, fed and watered correctly, exercised frequently, and treated with respect.
>
> It is only when a problem arises that we become conscious of our bodies. Then we respond by making corrections. The same thing happens in a factory. We should have a system in a factory that automatically responds when problems occur."

According to Ohno, TPS uses two pillars—"just-in-time" (synchronization) and "autonomation" (automation or machines that can prevent problems autonomously)—to eliminate waste and drive continuous improvement. Because there will always be a gap between actual and ideal synchronization, the process of approaching the ideal is an important aspect of lean operations. TPS, for example, *strives to make small but constant changes and improvements* (called ***kaizen***) by continuously identifying and eliminating sources of waste (i.e., by gradually lowering the water level to expose the rocks and then crushing them). We discuss this philosophy of continuous improvement further in Section 10.6. In this section, we focus primarily on concrete methods of lean operations to achieve synchronization and efficiency.

A lean operation has four ongoing objectives:

1. To improve process flows through efficient plant layout and fast and accurate flow of material and information
2. To increase process flexibility by reducing equipment changeover times and cross-functional training
3. To decrease process variability in flow rates, processing times, and quality
4. To minimize processing costs by eliminating non-value-adding activities such as transportation, inspection, and rework

The first three goals improve process synchronization, and the fourth improves cost efficiency. These goals are achieved through process-improvement levers discussed earlier in this book. Although the methods for achieving them will often be illustrated in the specific context of (automobile) manufacturing, the basic ideas work well in any stable, high-volume, limited-variety, sequential-processing environment including service industries.

The classic example of efficiency and synchronization for mass production was Henry Ford's Rouge, Michigan, plant in the 1910s. It was a totally integrated facility (including a steel mill and a glass factory), equipped with modern machine tools, electrical systems, and an automated assembly line and operated by highly paid, well-trained workers. Process efficiency was achieved by applying Frederick W. Taylor's principles of "scientific management," including time-and-motion studies, work rationalization, and best work methods (discussed in Chapter 2, Section 2.7). Streamlined to minimize the total flow time and cost, the moving assembly line was the ultimate in synchronizing production without buffer inventories between workstations. In fact, the roots of TPS can be traced to Henry Ford's system, except for one vital distinction, namely, the ability to handle product variety. Whereas Henry Ford's plant produced only the Model T (and only in black because the color dries fastest), modern automobile manufacturers must offer a wide variety of models and options, all of which must be of high quality and competitively priced, to satisfy contemporary customers' ever-rising expectations. We explore some of Toyota's tactics for meeting this challenge. However,

we keep our exposition at a more general level to address how a plant can achieve synchronization and efficiency through lean operations.

10.4.1 Improving Process Architecture: Cellular Layouts

A plant's process architecture (the network of activities and resources) has a significant impact on both the flow of work through the process and the ability of the process to synchronize production with demand. As indicated in Chapter 1, in a conventional functional layout, resources ("stations") that perform the same function are physically pooled together. Depending on their individual processing requirements, different product types follow different routings through these resource pools, and each flow unit may be sent to any available station in the pool.

A major advantage of the functional layout is that it pools all available capacity for each function, thereby permitting a fuller utilization of the resource pool in producing a variety of products. It also facilitates worker training and performance measurement in each well-defined function. Most important, it benefits from division of labor, specialization, and standardization of work within each function, thereby increasing the efficiency of each function. As we discussed in Chapter 1, a functional layout is ideal for job shops that process a wide variety of products in small volumes.

In terms of synchronization, however, the functional layout has several drawbacks. Flow units often travel significant distances between various resource pools, so their flow times are longer and it is harder to move them in small lots. The result is an intermittent jumbled flow with significant accumulation of inventories along the way. In addition, because each worker tends to be narrowly focused on performing only a part of the total processing task, he or she rarely sees the whole picture, leading to narrow, technology-focused process improvements.

An alternative to the process-based functional layout is the product-focused **cellular layout**, *in which all workstations that perform successive operations on a given product (or product family) are grouped together to form a "cell."* In order to facilitate a linear, efficient flow of both information and materials, different workstations within the cell are located next to one another and laid out sequentially. A cell is focused on a narrow range of customer needs and contains all resources required to meet these needs. Henry Ford's assembly line for the Model T is the classic example of such a product-focused layout. In a general hospital, trauma units, cancer care centers, and emergency rooms are examples of cells set up to process only patients with specific needs.

Advantages of Cellular Layouts The cellular layout facilitates synchronous flow of information and materials between processing stations. Physical proximity of stations within a cell reduces transportation of flow units between them and makes it feasible to move small batches (the ideal is one) of flow units quickly. It also facilitates communication among stations and improves synchronization by permitting each station to produce parts only if and when the following station needs them. Moreover, because any differences in workloads at different stations become immediately apparent, targeted improvements can be made to balance them. Similarly, if a station encounters a defective unit, that information can be reported to the supplier station immediately; because the supplier station has just handled the unit in question, the cause of the defect can be determined more easily. In short, the cellular layout facilitates synchronized flows and improved defect visibility, traceability, and accountability—which, in turn, leads to fast detection, analysis, and correction of quality problems.

Close interaction among different functions within a cell also encourages cross-functional skill development and teamwork among workers, which may lead to more satisfying jobs. Because the entire team works on the same product, workers can

experience a sense of ownership of the total product and process. The interchangeability of flexible workers also allows them to cooperate and smooth out any flow imbalances resulting from station variability. Finally, a cross-trained workforce improves synchronization by making it possible to adjust production volume to conform to changes in demand.

Disadvantages of Cellular Layouts Because resources are dedicated to specific cells, they cannot be used by other cells. Consequently, we lose the advantage of resource pooling that a functional layout enjoys. This loss of pooling can be countered with resources that are flexible and cross functional. Cells without flexible resources can be justified only if product volume is sufficiently high.

The stronger interdependence of cellular stations also means that worker incentives have to be based on team—rather than individual—performance. Because individual effort is only indirectly related to the team performance and rewards, workers have less incentive to do their share of work. One solution to this "free rider problem" relies on peer pressure to control the productivity of team members.

Thus, there are advantages and disadvantages to both functional and cellular layouts. Ideally, cellular structure is appropriate for products or product families with similar work-flow patterns and sufficiently high volume, as in automobile and electronic-goods manufacturing. In some cases it may be appropriate to set up a cell of very flexible resources that is assigned a large variety of low-volume parts. The focus of the cell is then on flexibility, and it produces all low-volume parts so that the rest of the plant can focus on producing the high-volume parts efficiently. If resources are not very flexible, it is inefficient to set up a cell to handle a variety of products with different work-flow requirements and high changeover times and costs, as in a job shop. Therefore, the functional layout is more appropriate.

10.4.2 Improving Information and Material Flow: Demand Pull

Given a system of interconnected stations in a processing network, managing flows means informing each station what to produce, when to produce, and how much to produce. There are two approaches to managing information and material flows: push and pull. *In the* **push** *approach, input availability triggers production,* the emphasis being on "keeping busy" to maximize resource utilization as long as there is work to be done. For example, using a popular planning tool called **material requirements planning (MRP)**, *the end-product demand forecasts are "exploded" backward to determine parts requirements at intermediate stations, based on the product structure ("bill of materials"), processing lead times, and levels of inventories at those stations.* A centralized production plan then tells each station when and how much to produce so that output will meet the planned (not the actual) requirements of downstream stations. In implementing the plan, each station processes whatever input quantity is on hand and pushes its output on to the next station. This push operation synchronizes supply with demand at each stage only under the following conditions:

- If all information (about the bill of materials, processing lead times, and parts inventories) is accurate
- If forecasts of finished goods are correct
- If there is no variability in processing times

Failure to meet any one of these conditions at any stage disturbs planned flow and destroys synchronization throughout the process, which then experiences excess inventories and/or shortages at various stages. Because each process bases output not on demand but on input availability, it is not surprising that production often fails to synchronize with demand.

An alternative method for ensuring synchronization is **pull**, where *demand from a customer station triggers production* so that each station produces only on demand from its customer station. Work at an upstream station is initiated by actual downstream demand from its customer station. Flow units are "pulled" from each process by its customer process only as they are needed rather than being "pushed" by the supplier process on to the customer process as they are produced. Under pull, the supplier does not produce or deliver anything until the customer really needs it and thus avoids inventories of unwanted outputs by refraining from processing inputs even if they are available.

Toyota's Taiichi Ohno (1988) characterized the pull system in terms of supermarket operations:

> From the supermarket, we got the idea of viewing the earlier process in a production line as a store. The later process (customer) goes to the earlier process (supermarket) to acquire the required parts (commodities) at the time and in the quantity needed. The earlier process immediately produces the quantity just taken (restocking the shelves).

The distinction between the push and pull systems of work flow is illustrated in Figure 10.3. Note that information that drives a push system is often a central plan based on the forecast of end-product demand. Information needed to determine flows in a pull system, in contrast, is local from the succeeding station only, and flows are controlled in a decentralized fashion.

There are two key requirements to making a pull system work:

1. Each process must have a well-defined customer, and each customer must have a well-defined supplier process.
2. Each process must produce the quantity needed only when signaled to do so by its customer.

Demand Signaling In a push system, input availability is sufficient to trigger production. In a pull system, however, the customer needs a signaling device with which to inform the supplier of its need.

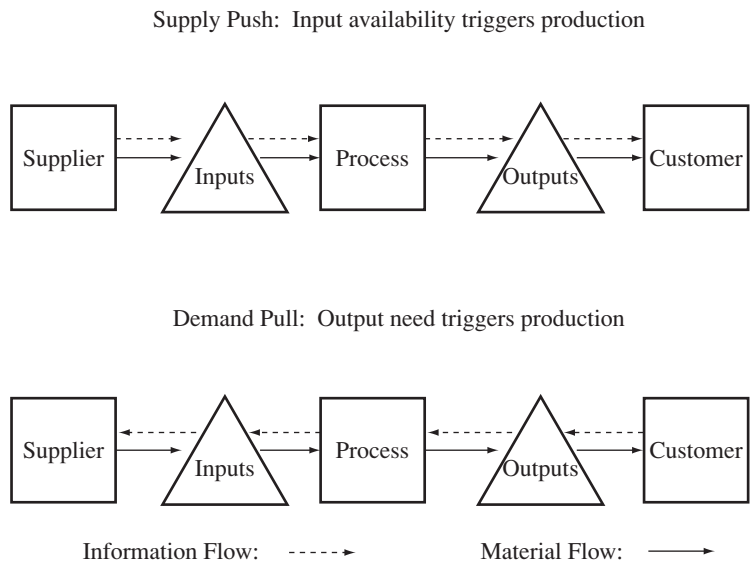


FIGURE 10.3 Synchronization: Supply Push versus Demand Pull

Toyota has formalized its signaling ***kanbans***, *a device that allows the customer to inform the supplier of its need. It is a card attached to an output flow unit in the buffer between customer and supplier processes and lists the customer process, the supplier process, parts description, and production quantity. Kanbans* are attached to output flow units in the buffer between customer and supplier processes, and each card lists the following information:

- Customer process
- Supplier process
- Parts description
- Production quantity

As the customer withdraws output flow units from the buffer, the attached *kanban* goes back to the supplier, which signals an authorization for the supplier to produce the listed quantity. On producing the stipulated quantity, the supplier returns the output with an attached *kanban* to the buffer. (There are actually two types of *kanbans*—one to authorize withdrawal and one to authorize production; however, here we will skim over the details.) Because each *kanban* corresponds to a fixed quantity of flow units to be produced and passed on, the number of *kanbans* in the buffer between the customer and the supplier determines the maximum size of the buffer. A station can produce a prescribed quantity only if it receives a production authorization *kanban*. Thus, *kanbans* control buffer inventory and provide information and discipline to the supplier as to when and how much to produce. The end customer's demand starts a chain reaction of withdrawals and replenishments of intermediate parts that ripples back through upstream stations. The EOQ-ROP system discussed in Chapters 6 and 7 can also be viewed as a pull system with the ROP (reorder point) triggering production at the supplier and the EOQ (economic order quantity) determining the quantity produced.

In the case of a process that handles multiple products, in addition to when and how much to produce, each supplier station must also know what to produce next. In an automobile assembly plant, for example, cars of different colors and options have different parts and processing requirements. A station that installs inner trim in 1 of 10 options needs to know which trim to install in the car next in line; likewise, its supplier needs a signal to produce that particular trim. One solution for handling variety is to create separate *kanbans* for each option—a system where 10 different buffers are controlled by 10 different *kanbans*. As the assembly station installs a particular trim, the released *kanban* signals its supplier to replenish that unit.

In order for the assembly station to know which trim unit to install on the car at hand, it needs to know the exact production sequence of cars rolling down the line. There is an alternative to maintaining multiple *kanbans* and complete information at each station if the trim supplier's response time is short enough to produce and deliver the trim to the assembly station in the period between when the production sequence is fixed and the time at which the car reaches the assembly station. Knowing the production sequence for the cars, the supplier can deliver different trims in the correct sequence. The assembly station can simply pick up the trim at the head of the buffer and install it into the next car without knowing the entire production sequence because the delivered trim sequence matches the car sequence coming down the line. In this case, only the trim supplier must know the production sequence to determine what to produce and in what sequence to deliver it. We refer to this approach of delivering parts in sequence as synchronized pull. This approach requires a greater capability on the part of the supplier and very tight coordination between the supplier and customer processes. At the same time, however, it achieves synchronization within a plant with minimal flow of material and information.

10.4.3 Improving Process Flexibility: Batch-Size Reduction

In addition to knowing what and when to produce, each station in a processing network needs to know how much to produce at a time. Consider, for example, an automobile assembly line that produces two different models—say, sedans and station wagons. Suppose that monthly demand for each model is 10,000 units. One way to meet this demand is to spend the first half of the month producing 10,000 sedans and the second half producing 10,000 station wagons. This pattern of production will not synchronize supply with demand because actual monthly demand is unlikely to look like this. Moreover, this approach places an uneven workload on the upstream processes (typically, suppliers) that feed parts for the two models: parts suppliers for station wagons have no orders in the first half of the month, and those for sedans have no orders in the second half of the month.

Level Production At the other extreme, we can achieve perfect synchronization if we alternate sedan and station wagon production one at a time. This results in **level production** (***heijunka*** in TPS terminology) *where small quantities are produced frequently to match with customer demand.* If monthly demand called for 10,000 sedans and 5,000 SUVs, a level production system calls for producing two sedans followed by one SUV and then repeating the sequence. If the demand pattern is stable, level production achieves perfect synchronization, producing flow units on demand and in the quantity demanded. Moreover, level production places an even workload on both the production process itself and all supplier processes feeding it.

Changeover Costs and Batch Reduction Level production in a multiproduct setting requires reducing the batch size produced of each product. As observed in Chapter 6, this reduction is economical only if the fixed cost associated with producing each batch can be reduced. The fixed cost results from the changeover cost and time required to switch production from one model to the other. Thus, a fundamental requirement of level production is reduction of changeover cost. Otherwise, excessive changeover costs from producing small quantities will drive up total production costs.

This concept of small-batch production is a focus for Toyota when introducing suppliers to lean operations. Changeover costs can be reduced by studying and simplifying the changeover process itself, using special tools to speed it up, customizing some machines, and keeping some extra machines that are already set up. All changeover activities that can be performed with the machine running (e.g., obtaining tools required for the changeover) should be completed without shutting down the machine. This reduces the time that a machine is not operating during the changeover, thus decreasing the changeover cost and time. The goal is to externalize as much of the setup as possible and perform these tasks in parallel with actual machine operation. By focusing on improvements to the changeover process itself, Toyota and other auto manufacturers have successfully reduced changeover times and costs by orders of magnitude. This increased ability to economically produce small batches without hurting the throughput results in low flow times and inventory.

The concept of small-batch production within a plant can be extended to small-batch pickups and deliveries made from several suppliers to several plants. One of two procedures is normally used: Either a single truck from one supplier carries deliveries to multiple plants, or a single truck destined for one plant carries small quantities of supplies from multiple suppliers. In either case, it is feasible to ship in smaller batches because the fixed cost of a shipment is spread over several suppliers or several plants.

We should reemphasize, however, that although level production is the goal of synchronization, it can be achieved economically only through reduction of the fixed setup (i.e., changeover) or transportation costs associated with each batch. Recall that

reduction of changeover cost was among the key levers explained in Chapter 6 and that it may not be optimal for every process to achieve level production with batches of one. In automobile manufacturing, for instance, expensive parts like seats are produced and delivered in batches of one. In contrast, windshield wipers, fasteners, and other low-cost items arrive in larger batches because it makes little economic sense to reduce batch sizes once the costs of doing so outweigh the benefits. In general, reducing batch size is most beneficial for large, expensive inputs; smaller, less expensive inputs are better handled in larger batches.

10.4.4 Quality at Source: Defect Prevention and Early Detection

Synchronization means more than just supplying correct quantities at correct times as required by customers: It also means meeting their quality requirements. Supplying defective flow units increases average flow time and cost because it necessitates inspection and rework. Moreover, in order to avoid starving the customer station, the production process must compensate for defective units by holding extra safety inventory in the buffer. In turn, this requirement further increases average flow time and cost. Thus, a key requirement of lean, synchronous operations is producing and passing only defect-free flow units between workstations. It requires planning and controlling quality at the source rather than after the fact (in final inspection) and can be accomplished in two ways:

1. By preventing defects from occurring in the first place
2. By detecting and correcting them as soon as they appear

Defect Prevention As discussed in Chapter 9, defect prevention requires careful design of both product and process. The goal is to use simplification, standardization, and mistake-proofing to minimize the chance of errors. Two techniques used by TPS to guard against defects are mistake-proofing (***poka yoke***) and intelligent automation ("**autonomation**" **or** ***jidoka***). Under *poka yoke*, for example, parts are designed to minimize chances of incorrect assembly. Under *jidoka*, machines are designed to halt automatically when there is a problem (deviation from the standard operating procedure). "Expanding this thought, we establish a rule that even in a manually operated production line, the workers themselves should push the stop button to halt production if any abnormality appears." (Ohno 1988, p. 7). Product and process design for defect prevention requires clearly defined and documented processing steps, thus removing worker discretion to the extent possible. Integrated design requires joint cooperation and input of all players: customers, designers, engineers, suppliers, as well as production workers. Each of them may have unique ideas and suggestions for product and process improvement that should be encouraged and rewarded.

Defect Visibility Even though all defects cannot be prevented, their early detection and correction is more effective and economical than catching them during final inspection. Early detection of defects improves our chances of tracing them to their sources. Early detection also reduces the waste of economic value that is added to flow units before their defects are caught and then must be reworked or discarded as scrap. Early detection contributes to better synchronization and lower costs in the long run by reducing the number of defectives introduced into the process stream.

Fast detection and correction of quality problems requires constant vigilance and monitoring of feedback. As discussed in Chapter 9, statistical process control can be used to monitor process performance so that any abnormal variation can be detected and eliminated early to maintain process stability.

Decentralized Control In addition, employees must be equipped with both the authority and the means to identify and correct problems at the local level without

administrative and bureaucratic delays. The main idea behind making problems visible is to minimize the cost and delay associated with identifying, searching for, and eliminating their sources. For example, in a Toyota plant, workers pull a rope conveniently located next to their stations if they detect a problem. Pulling the rope lights a *lamp on a signboard that immediately calls the supervisor's attention to the worker's location* (like a flight attendant's light in an airplane) called an ***andon***. The supervisor then rushes to the station to help correct the problem. If the supervisor is unable to correct the problem quickly, the line shuts down, alerting more people about the problem. The system is designed to increase attention to a problem until it is resolved. (One should consider the trade-off between the benefits of detecting and fixing problems early and the costs of lost production due to line stoppages.) Compare this practice to that of conventional plants in which resource utilization is given the top priority, work stoppage is permitted only on rare occasions, and only managers are empowered to take action. In fact, in a typical conventional plant, line workers do not feel the responsibility, motivation, or security to point out problems.

In summary, poor quality disturbs flow synchronization through a process. The basic strategy of lean operations, therefore, is threefold:

1. Preventing problems through better planning
2. Highlighting problems as soon as they occur
3. Delegating problem solving to the local level

The goal is to take permanent, corrective action immediately, minimizing future recurrences of problems, thus ensuring quality at source.

10.4.5 Reducing Processing Variability: Standardization of Work, Maintenance, and Safety Capacity

Variability in processing often results from imprecise specification of the work, equipment malfunction, and breakdown. The first step in reducing processing variability is to standardize work at each stage and specify it clearly. At a Toyota plant, each station has posted next to it a standardized work chart showing the flow time at the station, the sequence of activities performed, and the timing to perform them for each car processed. Green and red lines mark the beginning and end of each workstation, and a yellow line in between marks a point by which 70% of the work should be completed. The advantage of this standardization is threefold. First, the standardization reduces variability that arises from changing personnel. Second, the standardization reduces variability from one production cycle to the next. Finally, standardization makes it easier to identify sources of waste that can be eliminated. It is much harder to identify waste if the process itself is not clearly specified.

Given the vulnerability of a process operating without inventories to downstream equipment failure, planned preventive maintenance is an important prerequisite for synchronizing supply and demand. In fact, TPS calls for workers themselves to handle light maintenance of their equipment on an ongoing basis with more complete maintenance scheduled during off-hours.

It is impossible to eliminate all sources of variability; some are simply beyond the manager's control. For example, labor strikes, snowstorms, fires, and other acts of nature can disrupt supply deliveries. As discussed in Chapters 7 and 8, there are only two practical means of dealing with supply or demand variability if delays are to be avoided: carrying safety inventory or keeping some safety capacity. Although processes should consider the trade-off between carrying safety inventory and safety capacity, lean operations try to minimize carrying safety inventory because it increases flow time and jeopardizes synchronization. Consequently, a lean process must maintain some safety capacity as protection against variability.

Safety capacity may be in the form of extra machines, workers, or overtime. Toyota, for example, does not schedule production for all 24 hours in the day. The residual capacity is used as overtime if scheduled production is not completed by the end of the day. Ideally, safety capacity in the form of machines or workers should be flexible so that it can be used as needed.

10.4.6 Visibility of Performance

As discussed by Swank (2003), a fundamental principle of lean operations is to measure process performance from the customer's perspective. For example, when process performance is evaluated by measuring average time per call at a call center, customer-service representatives are eager to get the customers off the phone. However, this may lead to customers having to make repeat calls because their concerns were not fully resolved the first time. To measure process performance from the customer's perspective, it is more effective to measure the percentage of customers whose problems are resolved in a single call. This ensures that customer-service representatives are focused on resolving customer problems as opposed to doing their best to keep the calls as short as possible. Similarly, measuring internal flow time within the picking process at a mail-order warehouse is of little interest to a customer. The customer cares about the flow time from when she places an order to when it is delivered. Thus, all processes at the warehouse should be geared to reducing the overall flow time. It is important to ensure that goals at all levels of the organization are linked to each other. The metric used to measure a manager's performance should be directly affected by the metric used to measure the people working under her.

One of the most important principles of lean operations is that actual performance, along with expectations, be very visible for each work cell. In one of its lean initiatives, Wipro installed a Visual Control Board (VCB) to highlight the status of the software work in progress. "The project manager placed an A4 sheet of paper in a central location with each team member's name and daily assignments for the week. At the end of each day, each team member indicates what percentage of the work he has completed. The VCB not only provides a place for the project manager to receive an overall status report and a check for team member loading, but it also allowed him to identify potential problems sooner and then to provide targeted assistance as appropriate." (Staats and Upton, 2009, p. 14). The goal of this visibility is not to assign blame and punish low performers but to provide quick feedback for corrective action in case of a problem and to give teams an opportunity for celebrating success in case of high performance. The visibility of expectations and performance also clarifies for employees that they will be evaluated and rewarded for objective results that they can track themselves.

10.4.7 Managing Human Resources: Employee Involvement

Implementing synchronization within a plant requires cooperation, contribution, and commitment on the part of all employees to work in a tightly linked, highly interdependent environment. Managing human resources is therefore a critical factor in lean operations.

Behavioral studies since Elton Mayo's famous "Hawthorne experiments" at Western Electric in the 1940s have shown that if workers are involved in the decision-making processes that affect their jobs, they are better motivated to contribute substantially to productivity improvement. The key concept behind these theories of employee involvement is the recognition that workers have talents, education, and experience that can be harnessed to improve the process.

Worker participation in problem-solving and process-improvement efforts (as in "quality circles") is an important component of lean operations. Based on the premise

that the workers closest to the job have the most expertise to provide suggestions for improvement, the employee-involvement approach suggests involving workers in all phases of the improvement cycle. It also argues that employees possess the most current and accurate information about a given problem. It stands to reason, therefore, that providing them the necessary training and tools—and, just as important, "empowering" them with the authority and responsibility to make job-related decisions—is the fastest method of implementing decentralized control. The authority to stop production when a defect is detected at Toyota is an example of such an approach.

In lean operations, workers are cross-trained both to provide the company with flexible workers and to give workers greater variety through job rotation. In addition to their regular jobs, work teams in cells may also be authorized to perform certain managerial duties, such as work and vacation scheduling, material ordering, and even hiring new workers.

Worker participation in such initiatives requires that employees not only have basic skills and education but also are willing to learn multiple tasks, work in team environments, and be committed to the success of the entire process. Lean operations, therefore, places a great importance on the recruiting and training of workers.

10.4.8 Supplier Management: Partnerships

Outsourcing materials—that is, buying them from someone else rather than making them—provides a flexible alternative to vertical integration. In modern manufacturing, purchased materials not only account for a major portion of the product cost but also represent a major source of quality problems. With lean operations, reliable, on-time deliveries of defect-free parts assume critical importance. External suppliers, therefore, constitute an essential resource that impacts product cost, quality, and flow time.

A conventional approach to supplier management calls for selecting several suppliers, making them compete against one another on price alone, and then monitoring them closely to ensure that they do not neglect quality and timely delivery. It often leads to adversarial and even hostile relationships between the supplier and the manufacturer. Synchronizing flow becomes very difficult if the product is sourced from many suppliers. The lean approach to supplier management, in contrast, calls for choosing only a few capable, reliable suppliers with whom to cultivate cooperative, long-term relationships. The buyer works to make the suppliers an extension of the plant by sharing information and helping them improve their own processes through training, technical, and even economic assistance and by extending long-term contracts as incentives to induce cooperation in synchronizing flows of inputs with the plant requirements.

In terms of actual deliveries, the conventional approach seeks quantity discounts by purchasing in large volumes and tries to ensure quality through extensive inspection of incoming material. Lean operations, in contrast, involve processing without inventories or quality inspection. Plant synchronization with supplier requires that defect-free material be delivered frequently, in small batches, and directly to the point of usage. In turn, small, frequent, reliable deliveries require supplier proximity and simplified buying and accounts-payable procedures. They also require that the supplier's process be able to produce small quantities on demand—that the supplier's plant be synchronized with the buyer's. Ensuring quality at source (without the non-value-adding activity of inspection) requires supplier capability and commitment to producing quality parts. It also requires open communication between the buyer's plant and the supplier on such matters as product design changes and possible improvements.

Supplier management involves treating suppliers as partners—which is quite a change from the conventional approach that regards suppliers as outsiders not to be

trusted. Even now, firms often interpret lean operations to mean requiring suppliers to produce and hold the parts inventory and supply just-in-time to the plant so that the plant can operate without raw materials inventory. Such a policy amounts to simply pushing the plant's raw materials inventory back on to the suppliers. In fact, the goal of lean operations should be to reduce inventories in the total supply chain by synchronizing both the supplier's process and the buyer's process. It is thus critical to manage suppliers as a part of one's business, working with them closely to help them improve their quality, delivery, and cost so that they are able to meet the plant's requirements and remain economically viable.

In Brazil, supplier factories surround the Blue Macaw assembly plant of General Motors (Wilson, 2000). The factories produce instrument panels, seats, and other components for the Chevrolet Celta, a low-priced vehicle for the Brazilian market. The suppliers are responsible for the design and engineering of their components. Suppliers control a portion of the assembly line where they handle assembly of an entire system that has typically been designed and engineered by them. For example, Lear Corporation controls door assembly at Blue Macaw and installs locks, windows, and other components onto the door. Similar ideas have been used by the Ford Motor Company and Volkswagen in Brazil. The increased involvement of suppliers in the design, engineering, and assembly phase has led to significant savings.

In summary, lean operations aim to sustain continuous flow processing in an economical manner by implementing four closely related principles:

1. Synchronize material and information flows by means of cellular layouts and demand pull mechanisms (Sections 10.4.1 and 10.4.2)
2. Increase flexibility by means of fast changeovers that permit smaller batches to level production (Section 10.4.3)
3. Reduce variability by means of work standardization and improved supplier reliability and quality, coupled with safety capacity, preventive maintenance, and fast feedback and correction (Sections 10.4.4 and 10.4.5)
4. Decrease processing costs by improving quality and eliminating non-value-added activities such as transportation, inspection, and rework (Chapters 4 and 5)

Each of these principles are further facilitated by setting goals that are consistent with customer needs, establishing visibility of performance (Section 10.4.6), and making long-term investment in workers and suppliers (Sections 10.4.7 and 10.4.8) leading to successful lean operations. As many companies have realized, any efforts to implement just-in-time operations that ignore these prerequisites are sure to fail. For example, if a process has a high setup cost and a high degree of variability, it will be uneconomical and inefficient to operate without cycle or safety inventories (for further discussion, see Zipkin, 1991).

10.5 IMPROVING FLOWS IN A SUPPLY CHAIN

Producing and distributing goods to meet customer demand involves flows through a complex network of processes that include raw materials suppliers, finished-goods producers, wholesalers, distributors, and retailers. This entire value-adding network is the supply chain. Managing a supply chain involves storing and moving products and information along the entire network in order to make products available to customers when and where they are desired at the lowest possible cost. The goal of a supply chain is to synchronize flows throughout the network to meet customer demand most economically.

In the previous section, we discussed key issues in achieving synchronization and efficiency within a plant—which is just one node in the entire supply chain. The plant,

however, can be considered a network of processing stages through which raw materials, components, and subassemblies flow to emerge as finished products. Its structure, in other words, is similar to that of a supply chain. Therefore, the concepts we applied at the plant level are equally applicable in synchronizing flows in supply chains. There are, however, three special challenges in managing a supply chain:

1. **Scale magnification:** This implies that issues that arise within a plant (relating, for instance, to economies of scale, inventory levels, and flow times) are magnified in a supply chain. For example, flow times between nodes in a supply chain could be orders of magnitude larger than those between processes within a plant. Similarly, economies of scale in transporting goods from one node to another in a supply chain are much larger because of the geographical distances involved.
2. **Multiple decision makers:** Because different nodes in a supply chain may have separate ownership, each with its own objectives, the supply chain consists of multiple decision makers. Aligning incentives among different agents is much more difficult and results in suboptimization of nodes: Locally optimal decisions made at one node may not be globally optimal for the entire supply chain.
3. **Asymmetric information:** Each (independent) decision maker may possess only private information and lack the global information necessary to synchronize its flows with the rest of the supply chain. Thus, even if a decision maker wishes to act in the best interests of the chain, he or she may not be able to do so.

In this section, we discuss the consequences of unsynchronized flows in a supply chain, identify their root causes, and propose some measures for improving synchronization and efficiency in a supply chain. For a more detailed discussion of this topic, see Chopra and Meindl (2009).

10.5.1 Lack of Synchronization: The Bullwhip Effect

A supply chain can be analyzed in terms of product and information flows. Products primarily flow toward the customer, whereas information flows in both directions. Information regarding orders flows upstream towards the supplier, whereas information on prices and product availability flows downstream to the customer. Matching supply and demand involves synchronizing product flows with customer demand. The ability to synchronize is affected by information flows in a supply chain.

Consider a simple, linear supply chain that consists of a manufacturer, a distributor, a wholesaler, and a retailer. Customer demand at the retailer starts a chain reaction of orders upstream all the way back to the manufacturer. Figure 10.4 shows typical order patterns faced by each node in such a supply chain. Note that the retailer's orders to the wholesaler display greater variability than the end-consumer sales, the wholesaler's orders to its distributor show even more oscillation, and the distributor's orders to the manufacturer are most volatile. Thus, the pattern of orders received at upstream stages becomes increasingly more variable than consumption patterns at the retail end.

This *phenomenon of upstream variability magnification* is referred to as the **bullwhip effect** *and indicates lack of synchronization among supply chain members.* Even a slight disturbance in consumer sales sends back magnified oscillations, as does the flick of a bullwhip. In a perfectly synchronized supply chain, the order pattern at each stage would mimic the consumption pattern at the retail end. In a supply chain that is not synchronized, however, information flows are distorted, leading to inventory accumulation at some stages and shortages and delays at other stages. The bullwhip effect has been observed by firms in numerous industries, including Procter & Gamble (P&G) in consumer products, Hewlett-Packard in electronics, General Motors (GM) in automobiles, and Eli Lily in pharmaceuticals.

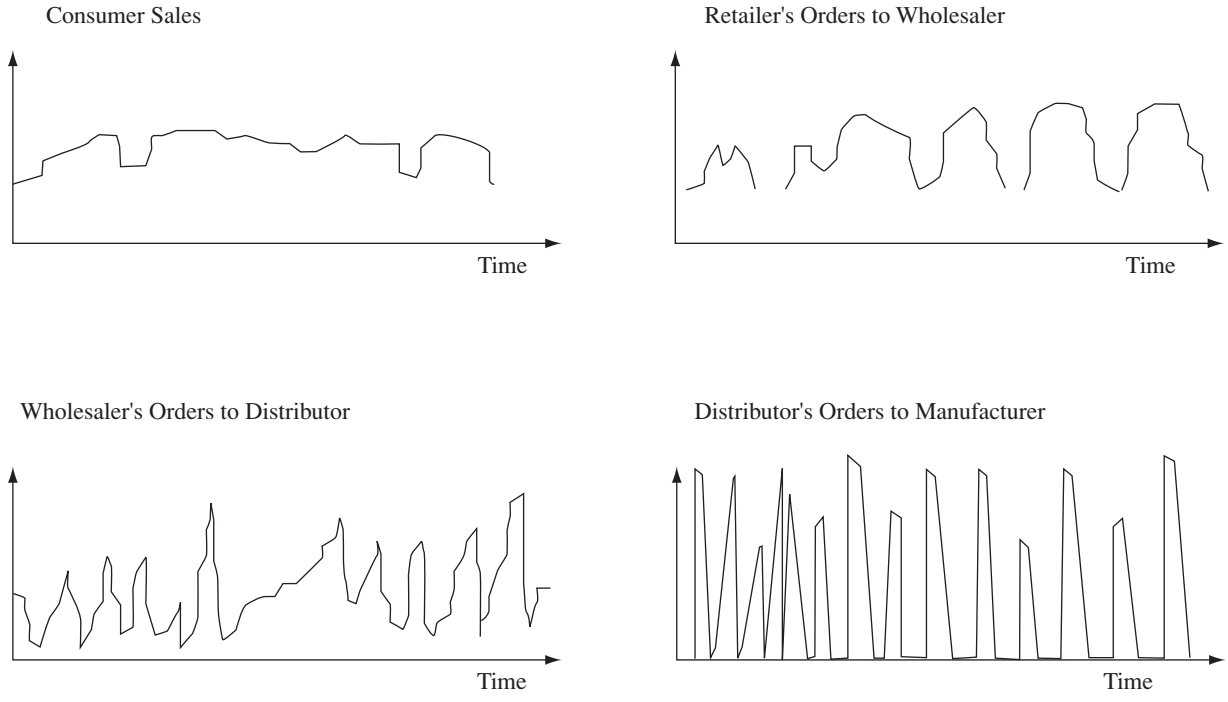


FIGURE 10.4 The Bullwhip Effect: Order Variablity in a Supply Chain

10.5.2 Causes of the Bullwhip Effect

Four main causes of the bullwhip effect have been identified by Lee et al. (1997):

1. Demand signal processing
2. Order batching
3. Price fluctuations
4. Rationing or shortage gaming

In the following sections, we discuss each cause briefly and show how it leads to an increase in the variability of the order pattern as compared to the demand pattern faced by each node in the supply network. The bullwhip effect is then a result of these four causes, as the variability cascades across the nodes as orders move upstream.

Demand Signal Processing Most firms rely on some form of demand forecasting to plan procurement, production, and capacity. Usually, short-term forecasting involves extrapolating the history of past sales and demand, with every observation of current demand factored into future demand projections. Consider a retailer who satisfies end-customer demand by ordering from a wholesaler. If demand in the current period is higher than expected, the retailer adjusts his forecast of the future leadtime demand (including both the mean and the error). The new forecast is adjusted upward because current realized demand is higher than the previous forecast. The retailer's order with the wholesaler increases because of the higher forecast of mean leadtime demand. The increase in order size is exacerbated by the delay in material and information flow between the two stages. The retailer's order is higher-than-realized demand to compensate for the delay in replenishment. The same phenomenon recurs when the wholesaler receives the retailer's order: The wholesaler's order to his or her supplier is also amplified relative to the retailer's order (whose order was amplified relative to the change in customer demand). Order amplifications cascade upstream in the supply chain because each stage forecasts demand based on orders it receives.

Conversely, if the current demand is lower than the forecast amount, the retailer adjusts the forecast of leadtime demand downward. This adjustment leads to a reduction in order, thus creating a distortion in the ordering pattern. Observe that the order amplification is exacerbated by the fact that each stage in the supply chain makes plans according to a different information set—namely, the order stream from the immediately downstream stage—and not according to the ultimate customer demand. Thus, we see an interplay of all three difficulties discussed earlier—scale magnification, diverse decision makers, and private information.

Order Batching The practice of batching occurs when a node in the supply chain places large and infrequent orders. Firms may place orders in some periodic fashion—say, on a weekly, biweekly, or monthly basis. This practice results from some form of economy of scale in procurement, production, or transportation, as discussed in Chapter 6. Firms also place orders in large batches in response to incentives (e.g., quantity discounts) offered by a supplier.

Although batching may be optimal for the buyer, it creates a distortion in the demand pattern experienced by the supplier. As discussed in Chapter 6, when demand rates are known and constant, the use of an economic order quantity (EOQ) model creates an order pattern with large spikes. If a supplier expects this pattern of ordering from a buyer (perhaps through information sharing), he or she can account for the unevenness. Often, however, this is not the case because only orders and not demand information are passed along the supply chain.

Whenever a process at one stage in the supply chain places orders in batches, the process at the upstream stage sees orders that are much more variable than end-customer demand. This effect is exacerbated when multiple retailers place large orders simultaneously with the same upstream supplier (numerous retailers, for instance, may place their orders every Monday).

Price Fluctuations When prices offered by an upstream stage to a downstream stage fluctuate often, the downstream stage may order more than it immediately needs when prices are low and postpone purchases when they are high. In Chapter 6, we showed how buyers forward buy and increase order quantities by a large amount when suppliers offer small short-term price discounts. Forward buying makes orders even more variable than demand, thus exacerbating the bullwhip effect. Short-term price discounts and forward buying are fairly common for several commodity products, such as dry goods in the grocery industry.

Rationing or Shortage Gaming When total orders placed by retailers exceed product availability, manufacturers use some form of *rationing* to allocate their products to buyers. If retailers know that a product will be in short supply (and thus rationed), they may exaggerate their needs when placing orders. When demand turns out to be lower than the inflated order quantities, retailers start canceling orders—leaving large levels of inventories with manufacturers. This pattern can set in even if shortages are not real: Orders may be amplified at the slightest *perception* of a shortage by retailers. Because orders do not reflect actual consumer demand, such "gaming" behavior on the part of downstream stages produces the bullwhip effect.

In November 2003, *Off the Record Research* reported wide shortages of Nokia phones in Europe. Buyers were quoted as saying, "We're seeing a lot of supply problems affecting almost all vendors at the moment. When I order 50,000, I'll probably get 20,000 from Nokia; it's the same with Samsung and Siemens." The result of widespread shortages was double ordering by network operators hoping to increase available supply. The double ordering was expected to affect inventory levels in early 2004, when the additional units were delivered.

10.5.3 Levers to Counteract the Bullwhip Effect

A typical supply chain is characterized by independent players who optimize their own objectives according to limited private information. As we have seen thus far, even when these players behave rationally, information distortion in the supply chain can produce the bullwhip effect. These root cause suggest the following levers for counteracting the bullwhip effect:

1. Inefficient processes (resulting, for example, in long flow times between stages or in high fixed costs)
2. Inconsistency of available information (due to poor timing, inaccuracy)
3. Local actions by individual players that are suboptimal for the overall system

Having understood some of the causes of the bullwhip effect, we now outline some levers to counteract them.

1. Operational effectiveness
2. Information sharing
3. Channel alignment

Operational Effectiveness Throughout this book, we have considered operational effectiveness in terms of cost, quality, and response time and suggested levers to achieve effectiveness in these terms. Several of these levers also help counter the bullwhip effect as outlined here:

- **Reduce (material and information) flow time:** The bullwhip effect is reduced if material and information flow times are decreased. Some technologies, such as electronic data interchange (EDI) and the Internet, permit various stages in the supply chain to transmit information electronically, thereby reducing delays in information flows. Cross-docking, which is widely practiced by Walmart and many other firms (see Example 2.1 in Chapter 2), calls for moving material directly from receiving to shipping with minimum dwell time in the warehouse—a practice that helps decrease the transportation flow time and pipeline inventory between suppliers and retailers.
- **Reduce economies of scale:** The bullwhip effect can be diminished if batch sizes of purchases are reduced. The various levers for decreasing batch size discussed earlier can be applied to reduce batch sizes in a supply chain:

  Reduce fixed costs: Fixed procurement, production, and transportation costs create the bullwhip effect by encouraging large batch order sizes. EDI and the Internet reduce fixed procurement costs by allowing firms to place orders electronically. Several principles that we have attributed to lean operations reduce changeover cost and encourage production in smaller batches. For example, **single minute exchange of dies (SMED)** *is a system by which the changeover times can be reduced to less than 10 minutes,* and **flexible manufacturing systems (FMS)** *is a reprogrammable manufacturing system capable of producing a large variety of parts.* Both permit level production (*heijunka*) by reducing production setup and changeover costs.

  Give quantity discounts for assortments: Suppliers often offer quantity discounts based on the batch size of purchase. These discounts, however, are typically offered for individual items. For example, a firm may offer a 3 percent discount for purchases in a full truckload. When these discounts are offered separately for each product family, customers have an incentive to purchase full truckloads of each family. The result is a distortion of ultimate demand information. If suppliers offer discounts on assortments of items, thus allowing the customer to obtain the same 3 percent discount as long as they fill a truck-

load of the assortment, there is little need to exaggerate batch sizes for individual items. Such a policy reduces distortion in item-level demand information while still permitting the supplier to exploit economies of scale in transportation. Following this approach, P&G now offers discounts to distributors as long as they fill a truckload with an assortment of P&G products.

Form logistical alliances: Another way to exploit transportation economies is to form an alliance with a third-party logistics firm. Such providers achieve economies of scale in transportation by consolidating the needs of multiple suppliers/customers. Consolidating shipments lessens the need to increase batch sizes by allowing each supplier/customer to ship/receive less than a full truckload of quantities. Firms should, however, consider other coordination and strategic issues before deciding to outsource the logistics function.

Information Sharing The presence of multiple decision makers working with private information affects the product/information flows in the supply chain. Sharing of information among supply chain members can reduce the magnitude of the bullwhip effect:

- **Share consumption information with upstream players:** Each stage in the supply chain processes its demand information to construct a forecast for the future (a strategy that we have labeled "demand signal processing"). However, only the last stage in the chain is privy to sales data regarding the end-consumer demand, which is usually collected through point-of-sale (POS) technology. Forecasts at all other stages are based on the orders they receive. Consequently, each stage in the chain is trying to forecast demand based on a different set of data. A first step in information sharing is to make sales data available to all players in the supply chain so that every member's plans are based on the same data set. In fact, as described in Chapter 2, Walmart shares its sales data with suppliers.
- **Share availability information with downstream players:** Shortage gaming results when retailers do not know the actual availability or capacity of their suppliers. Although sharing capacity/availability information will eliminate mistaken perceptions of shortages, it will also reveal the existence of real shortages. Thus, it may not be a perfect instrument to counteract the bullwhip effect. When shortages do exist, allocation policies should be based on past sales and not current orders.

Channel Alignment Although operational improvements and information sharing may assist independent supply chain players in making decisions that improve their own performance, these practices alone are usually insufficient to synchronize the entire supply chain. Other explicit coordination/incentive mechanisms are needed to align the priorities of individual members with those of the system:

- **Coordinate replenishment and forecasting decisions:** Even if every stage in the supply chain possesses customer sales data, differences in forecasting methods and buying practices can still lead to fluctuations in orders. One solution is for a single upstream stage to control the replenishment of material to the downstream stage. This tactic works when the upstream stage has access to downstream demand and inventory information and replenishes its stock accordingly. **Vendor managed inventory (VMI)** and **Continuous Replenishment Program (CRP)** are two techniques used by the consumer-products industry to implement these practices. VMI *is a partnership program under which the supplier decides the inventory levels to be maintained at its customer locations and arranges for replenishments to maintain these levels.* P&G and Walmart have been at the forefront of VMI programs initiated in 1985. Under CRP, *the supplier automatically replenishes its customer inventories based on contractually agreed-on levels.* The Campbell Soup Company uses a CRP program with its retailers to coordinate replenishment.

Under such programs, however, a downstream stage, which no longer decides on how much to order, may perceive a loss of control. Another solution, therefore, is to adopt a coordinated forecasting and replenishment methodology for all stages in the chain. **Collaborative Planning, Forecasting, and Replenishment (CPFR)** *is an initiative in the consumer-goods industry designed to coordinate planning, forecasting, and replenishment across the supply chain.* Details on CPFR can be found at www.cpfr.org.

- **Stabilize prices:** Short-term price reductions provide an incentive to the retailers to "forward buy" and thereby distort the supplier's demand information. A manufacturer can reduce forward buying by the following methods:
  1. Establish a uniform wholesale-pricing policy
  2. Limit the amount that can be purchased under forward buys
  3. Credit retailers for promotional discounts based on customer sales during a given period rather than on orders placed

  In the grocery industry, for instance, major manufacturers like P&G and Kraft have adopted everyday low purchase pricing (EDLPP) strategies, a practice discussed in Chapter 6.
- **Change allocation policies:** We have already observed that sharing upstream capacity/availability information is not a perfect instrument for reducing the bullwhip effect due to gaming. Allocation mechanisms based on current orders encourage downstream stages to exaggerate their orders. However, other allocation mechanisms—such as GM's policy of basing allocations on past sales—may remove incentives to inflate orders.

To summarize, the ability to synchronize flows in a supply chain is affected by such factors as operational efficiency, information availability, and level of coordination. Organizations must understand the root causes of the inefficiency that results from the bullwhip effect and take measures to remedy them. Implementation of the solutions like those proposed in this chapter is challenging because the process often involves interorganizational issues in coordination, information sharing, and change of incentive structures.

10.6 THE IMPROVEMENT PROCESS

Although we have identified various levers for process improvement (i.e., what to do), we have not yet described any process or framework for achieving such improvement (i.e., how to do it). In this section, we conclude by discussing the process of process improvement. We describe an improvement process that begins by maintaining process stability in the short run, gradually improving it over time, and occasionally revamping it in the long run. Managers should keep in mind that the workers closest to the job are in the best position to identify opportunities for process improvement. They should, therefore, be encouraged—even prodded—to make suggestions. Including workers in the improvement process makes them feel not only comfortable and secure in exposing and facing up to new problems but also disciplined and ambitious enough to overcome complacency in approaching the ideal.

10.6.1 Process Stabilization: Standardizing and Controlling the Process

The first step in process improvement is to define the process by standardizing the various activities. Because process performance usually displays variability, reliable measurement requires that the observed performance be statistically predictable. The second step in process improvement is to bring the process under control, as discussed in Chapter 9. Statistical process control involves monitoring performance over time and providing fast feedback when variability appears to be abnormal. This practice identifies

and corrects sources of abnormal variation, thus ensuring that our estimates of its performance characteristics are reliable. At this stage, we have a well-defined process and a reliable measurement of its performance in relation to the ideal. This is a base from where improvement should start.

10.6.2 Continuous Improvement: Management by Sight and Stress

For a process that has been stabilized, any gap between actual and ideal performance can be reduced by making **continuous improvement**, *ongoing incremental improvement in process performance,* which is an important aspect of lean operations and TPS. As we mentioned in Section 10.4, it is often referred to as *kaizen,* which means "a good change." *Kaizen* has also been characterized as "continuous, incremental improvement in the process by everyone involved" (Imai, 1986). We next discuss three drivers of continuous improvement.

Management by Sight Management by sight focuses on driving continuous improvement by making problems and opportunities visible and providing incentives and tools to eliminate the former and take advantage of the latter. A natural short-term tendency in process management is to cover up process imperfections (such as inflexibility or variability) by building in safety cushions (such as cycle or safety inventory). This approach obstructs our view of process imperfections and reduces our sense of urgency to remove them. The principle of management by sight calls for constantly removing—not inserting—safety cushions such as inventory in order to expose and contend with process problems (as opposed to covering up these problems). Both removing safety cushions and exposing problems are treated as an opportunity to improve the process.

Management by Stress Management by stress focuses on constantly stressing the system, thus forcing it to improve performance to reduce the stress. In the river analogy, as the water level is lowered, new problems surface, and we are forced to deal with them. As soon as we have solved these newly visible problems, the water level is lowered again, exposing yet more rocks to be dealt with. Our goal is to refuse to be content with smooth sailing in ample water. Rather than using extra resources as a safety cushion to protect against imperfections, we keep on reducing their level, relentlessly exposing more problems, and eliminating their root causes. The idea is to keep the process under constant pressure by gradually removing the security blanket.

The management-by-stress philosophy teaches us that by keeping the system under constant stress, we force new problems to become visible and so increase our own responsibility for eliminating root causes rather than simply reacting to them as they occur. At Toyota, for example, the rope-pull (*andon*) system is a tool for making problems visible. If a problem repeatedly prompts rope pulls, it is clearly in the supervisor's best interest to get to the root cause of the problem and eliminate it once and for all. To ensure sufficient visibility, Toyota tracks the number of rope pulls per shift. Similarly, the production process is kept under constant stress by removing *kanbans* (and hence the inventory buffer) between stages so that stages or nodes are forced to devise methods to work with less and less inventory.

Recall, however, that the success of continuous improvement requires a gradual lowering of the water level; otherwise, the boat will scrape the bottom against a rock and spring a leak. In fact, the failure to appreciate the importance of gradual stress is one reason why lean operations sometimes fail in practice: Firms set arbitrarily high targets for inventory reduction without correspondingly improving the process. Ultimately, it is process improvement—not merely inventory reduction in itself—that is the goal. Inventory reduction is merely a means for exposing problems.

Management by Objective Another approach to continuous improvement is to regularly set targets (say, every quarter or every year) for critical performance measures.

These targets should reflect both the demands of the marketplace and the performance of competitors. The purpose of such targets is to guide the firm in its efforts to do better than competitors with the same level of resources or to do as well with fewer resources. In either case, targets, once achieved, are then revised, and the process (and the pressure) continues. This approach to continuous improvement is sometimes called "management by objective." The major difference between this approach and that of management by sight and stress is a matter of focus. In managing by sight and stress, we focus on making problems visible and then treat them as drivers for change. In management by objective, we focus on achieving an objective and let the desire to hit targets drive change. In either case, the basic tool kit for process improvement contains the same levers that we have been describing throughout this book.

10.6.3 Business Process Reengineering: Process Innovation

Sometimes, gradual improvement toward the ideal process may not be enough: A significant jump in process performance may be necessary to catch up to or overtake the competition. In that case, the solution might be **reengineering**, which Hammer and Champy (1993) (who popularized the term in the early 1990s) define as "*fundamental rethinking and radical redesign of business processes in order to achieve dramatic improvements in critical contemporary measures of performance such as cost, quality, service and speed*."

Consider some of the key terms of this definition:

- Fundamental rethinking means reexamining and questioning traditional practices and assumptions that may have become obsolete, erroneous, or inappropriate over time. At each stage of a business process, one asks why it is done and how it can be done differently or, better yet, how it can be eliminated completely.
- Radical redesign means reinventing the process from scratch, not just tweaking the existing one. It means channeling a new river without rocks rather than trying to uncover and remove the rocks one by one in the current river. It thus requires starting with a "clean slate" and asking what an ideal process would look like if we had to start all over again.
- Dramatic improvements mean substantial changes aimed at, say, 10-fold, not 10 percent, improvements in performance measures. For example, the goal of a typical reengineering project would be to design, produce, and deliver a product with half as many defectives, in half as much time, at half as much cost as the one we now market. The six-sigma philosophy that we discussed in Chapter 9 is an example of setting such "stretch goals." Improvements of such magnitude require "out-of-the-box" innovative thinking.
- Critical measures of performance that are important to the customer—cost, quality, and response time—should be the focus of process improvement. It is a waste of energy to make improvements that the customer does not see or value.

Reengineering versus Continuous Improvement As a framework for improvement, reengineering differs from continuous improvement in three elements: magnitude and time frame of desired improvement and change drivers. In continuous improvement, the change drivers (visibility or targets) are internal components of the existing process. In process reengineering, a complete rethinking of the process itself is forced by something external—either a dramatic change in customer needs or a change in technology. We strive not merely to make the existing process better but, potentially, to invent a new process that will do the job significantly better than the current process.

Hammer and Champy (1993) cite the accounts payable process at Ford Motor Company as a classic example of successful reengineering. During the early 1980s, the department employed over 500 people in North America. Management hoped for a 20 percent reduction in head count by improving the existing processes. That process,

however, focused on processing invoices sent in by suppliers, and Ford discovered that by eliminating invoices altogether and basing payments on receipts of shipments, it could radically alter the entire procurement process. The result was a 75 percent reduction in the personnel. Note, however, that although the popular press tended to equate "reengineering" with "downsizing" in the early 1990s, reducing head count and cost are not the only projects in which reengineering may be useful. Reengineering may also make dramatic improvements in terms of time, quality, and flexibility. Hammer and Champy discuss several illuminating examples.

Reengineering and continuous improvement are not necessarily antithetical approaches. Both are valuable as components of the same long-term improvement program. In fact, along with the design for a new process, reengineering effort should also include a program for continuous improvement. Similarly, while continuous improvement takes a process toward ideal performance in regular, incremental steps, reengineering is needed from time to time to make a radical change—especially when significant changes occur in the external environment and technology. Thus, reengineering is called for when there is a dramatic change in customer expectations or a change in technology that makes possible a completely different process design.

10.6.4 Benchmarking: Heeding the Voices of the Best

Process improvement requires setting and approaching specific goals—a project that can be aided greatly by studying others' processes and emulating their best practices; it can save your time and money by not having to "reinvent the wheel." Robert Camp (1995) defines **benchmarking** as "*the process of continually searching for the best methods, practices, and processes, and adopting or adapting the good features to become 'the best of the best.'*" We may benchmark someone else's products (in terms of price, quality, or response time), key processes (in terms of cost, flow time, or inventory turns), or support processes (such as warehousing, billing, or shipping).

In search of best practices, we may look either within our own organization or to outside competitors. We may even turn to noncompetitors in other industries. We have already seen, for instance, how the Japanese devised the pull system of material flow based on observations of U.S. supermarket operations. Xerox Corporation benchmarked mail-order retailer L.L. Bean for its efficient logistics and distribution system. In an effort to expedite aircraft turnaround times at the gate, Southwest Airlines studied the National Association for Stock Car Auto Racing (NASCAR) pit stops. The use of bar coding, so prevalent at supermarket checkout counters, is now widely used by manufacturers to manage parts inventories and flows.

The key to successful benchmarking is not merely duplicating the activities of others: Benchmarking means identifying the basic concepts underlying what world-class companies do, understanding how they do it, and adapting what we have learned to our own situation. It requires external orientation to identify the best in class, openmindedness to understand their approach, and innovativeness to modify their solution to fit our problem.

10.6.5 Managing Change

Process improvement means changing our way of doing business, which is accompanied by uncertainty. It is a natural human tendency, however, to prefer the status quo and predictability. In managing change, the challenge is to encourage people to accept change and to motivate them to take the kinds of risks that bring about change for the better.

Ironically, it is easier to motivate people to change when times are bad; change is then perceived as a survival imperative rather than an option to improve. By that time, however,

it may be too late to improve the firm's competitive position merely by making gradual improvements to the existing process; it may be necessary to reengineer the whole process. It is also unfortunate that when times are good, the natural tendency is to be complacent—and perhaps oblivious to potential environmental changes and competitive threats. The challenge then is to make people feel dissatisfied enough with the status quo to seek change and yet feel secure enough to take risks associated with a change. As we saw in our discussion of continuous improvement, this motivational balance can be spurred by increasing visibility of waste (management by sight), gradually reducing available resources (management by stress), or gradually raising goals (management by objective).

Finally, any organizational change is easier to bring about if everyone affected by it is involved in a participatory spirit, in a nonthreatening environment, with open lines of communication.

Summary

In this chapter we examined problems of managing flows of multiple products through a network of processes. The overall goal of such a processing network is to match its supply with demand in terms of cost, quality, variety, and availability of products when and where customers want them. The ideal is to achieve this synchronization of supply and demand at the lowest cost. We studied principles of lean operations to accomplish this at the plant level and extended them to managing the entire supply chain. The objective is to eliminate waste of all kinds excess costs, defects, delays, and inventories by increasing process capability and flexibility and decreasing variability. Furthermore, different processes must be coordinated to facilitate fast and accurate flow of information and materials between them.

Concrete methods for accomplishing these objectives at the plant level include the following:

1. Improve process architecture with a cellular layout
2. Coordinate information and material flow using a demand pull system
3. Decrease batch sizes and improve process flexibility by reducing changeover times
4. Ensure quality at source
5. Reduce processing variability by work standardization, preventive maintenance, and safety capacity
6. Increase visibility of performance to identify areas for improvement
7. Involve all employees in the improvement process
8. Coordinate information and align incentives with suppliers

Similar principles apply in managing flows in a supply chain, with added complications due to scale magnification and multiple decision makers having different information and incentives. This results in an increased variability in orders due to divergent forecasts, order batching, price fluctuations, and shortage gaming. Levers to synchronize the supply chain include the following:

1. Reduce information and material flow times through technology and logistics
2. Decrease fixed costs of production and ordering and reduce quantity discounts
3. Share information on customer demand and product availability
4. Coordinate forecasts and replenishment between various parties
5. Avoid short-term price fluctuations

Improvement in the process at the plant or supply chain level requires (1) stabilization through process standardization and control; (2) continuous improvement through management by sight, stress, and objectives; and (3) process reengineering through completely rethinking and radically redesigning the process for dramatic improvements in critical performance measures such as cost, quality, and response time that are important to the customers. Benchmarking accomplishments of the best in class in these areas helps set and achieve concrete goals for process improvement by identifying, adapting, and adopting their practices. Finally, it is important to recognize that every process-improvement effort entails an organizational change that must be motivated and internalized.

Key Terms

- *Andon*
- Autonomation
- Benchmarking
- Bullwhip effect
- Cellular layout
- Collaborative Planning, Forecasting, and Replenishment (CPFR)
- Continuous improvement
- Continuous Replenishment Program (CRP)
- Flexible manufacturing systems (FMS)
- *Heijunka*
- Ideal process
- *Jidoka*
- Just-in-time (JIT)
- *Kaizen*
- *Kanbans*
- Level production
- Material requirements planning (MRP)
- Plant
- *Poka yoke*
- Process efficiency
- Process synchronization
- Processing network
- Pull
- Push
- Reengineering
- Single minute exchange of dies (SMED)
- Supply chain
- Vendor managed inventory (VMI)
- Waste

Discussion Questions

10.1 How does the use of kanbans result in a pull system?

10.2 A manufacturer of auto parts has just learned about the Toyota Production System and is trying to implement lean operations. Traditionally, there has been no control on the amount of work-in-process inventory between stages (it has been known to exceed 500 parts between some stages). As a first step, the amount of work-in-process inventory between stages is limited to a maximum of 20 parts. What, if any, impact does this have on the output from the factory in the short term? Using lessons learned from the river analogy, how should the manufacturer manage buffers?

10.3 Taiichi Ohno, architect of the Toyota Production System, claims to have been inspired by U.S. supermarkets in his use of kanban cards. Describe how a supermarket dairy section fits in with use of kanbans in the Toyota Production System.

10.4 List conditions under which a cellular layout is most beneficial. Under what conditions is a functional layout to be preferred?

10.5 What are some mechanisms to implement a pull system in a multiproduct plant? What are the pros and cons of each mechanism?

10.6 What are some advantages of *heijunka*, or level production? Why are short changeover times essential if *heijunka* is to succeed?

10.7 At each stage of a supply chain, why do forecasts based on orders received lead to the bullwhip effect, especially if lead times are long? What countermeasures can improve the situation?

10.8 Why do price promotions exacerbate the bullwhip effect? What countermeasures can improve the situation?

10.9 What actions can a firm like Walmart take to help diminish the bullwhip effect on its supply chain?

10.10 What actions can a firm like P&G take to help diminish the bullwhip effect in its supply chain?

Selected Bibliography

Camp, R. *Business Process Benchmarking.* Milwaukee: ASQ Quality Press, 1995.

Chopra, S., and P. Meindl. *Supply Chain Management: Strategy, Planning, and Operation, 4th ed.* Upper Saddle River, N.J.: Prentice Hall, 2009.

Doig, S. J., A. Howard, and R. C. Ritter. "The Hidden Value in Airline Operations." *McKinsey Quarterly,* no. 4 (2003): 105–115.

Gualandi, G., R. Hatfield, J. King, H. Leuschner, P. Ridgewell, and G. Shao. "Nokia Feels the Squeeze from Shortage." *Off the Record Research: Sound Byte,* November 13, 2003.

Hammer, M., and J. Champy. *Reengineering the Corporation: A Manifesto for Business Revolution.* New York: Harper Press, 1993.

Hopp, William J., and Mark Spearman. *Factory Physics: Foundations of Manufacturing Management,* 2nd ed. Burr Ridge, IL: Irwin, 2004.

Imai, M. *Kaizen: The Key to Japan's Competitive Success.* New York: Random House, 1986.

Lee, H. L., V. Padmanabhan, and S. Whang. "Information Distortion in a Supply Chain: The Bullwhip Effect." *Management Science* 43, no. 4 (April 1997): 546–558.

Mishina, Kazuhiro. *Toyota Motor Manufacturing, U.S.A., Inc.* Harvard Business School Case Study 1-693-019. Boston, MA: Harvard Business School, 1992. 1–23.

Ohno, T. *Toyota Production System: Beyond Large-Scale Production.* Cambridge, Mass.: Productivity Press, 1988.

Passariello, C. "Louis Vuitton Tries Modern Methods on Factory Lines." *Wall Street Journal,* October 9, 2006. Page A1.

Pollack, A. "Aerospace Gets Japan's Message: Without Military Largess, Industry Takes the Lean Path." *New York Times,* March 9, 1999.

Schoenberger, R. *Japanese Manufacturing Techniques: Nine Hidden Lessons in Simplicity.* New York: Free Press, 1982.

Spear, Steven, and H. Kent Bowen. "Decoding the DNA of the Toyota Production System." *Harvard Business Review* 77, no. 5 (September–October 1999): 97–106.

Staats, B. R., and Upton, D. M. "Lean Principles, Learning, and Software Production: Evidence from Indian Software Services," Working Paper 08-011, Harvard Business School, 2009.

Swank, C. K. "The Lean Service Machine." *Harvard Business Review* 81, no. 10 (October 2003): 1–8.

Wilson, A. "Blue Macaw: GM's Factory of the Future." *Automotive News International,* September 1, 2000.

Womack, J. P., D. Jones, and D. Roos. *The Machine That Changed the World: The Story of Lean Production.* New York: Macmillan, 1990.

Wysocki, B., Jr. "To Fix Healthcare, Hospitals Take Tips from Factory Floor." *Wall Street Journal,* April 9, 2004.

Zipkin, P. "Does Manufacturing Need a JIT Revolution?" *Harvard Business Review* 69, no. 1 (January–February 1991): 40–50.

This page intentionally left blank

APPENDIX I

MBPF Checklist

Here we provide a summary of key points from the book. This appendix is meant to serve as a checklist for managing business process flows.

PROCESS FLOW MEASURES

- ***Key concepts:*** Flow time (T), inventory (I), throughput (R), process cost (c), quality
- ***Key relation:*** Inventory = Throughput × Flow time: $I = R \times T$
- ***Key management activity:*** Select process flow measures to manage for improvement
- ***Key metrics:*** Net present value, return on total assets

Because the three operational measures (flow time, inventory, and throughput) are interrelated, defining targets on any two of them defines a target for the third. The basic managerial levers for process improvement are the following:

1. Increase in throughput (decrease in flow time)
2. Decrease in inventory (decrease in flow time)
3. Decrease in process cost
4. Improvement in process quality

LEVERS FOR MANAGING THEORETICAL FLOW TIME

- ***Key concepts:*** Critical path, critical activity, theoretical flow time
- ***Key management activity:*** Identify and manage activities on all critical paths
- ***Key metric:*** Length of critical paths

Because the theoretical flow time of a process is determined by the total work content of its critical path(s), the only way to decrease it is by shortening the length of every critical path. The basic approaches to decreasing the work content of a critical path are the following:

1. Move work content off the critical path ("work in parallel").
2. Eliminate non-value-adding activities ("work smarter")
3. Reduce the amount of rework ("do it right the first time")
4. Modify the product mix ("do the quickest things first")
5. Increase the speed of operation ("work faster")

In addition, the flow time of a process is impacted by waiting time. The levers for reducing waiting time include:

- Managing the effects of congestion
- Reducing batch sizes
- Reducing safety buffers
- Synchronizing flows

These levers are discussed in Chapters 5, 6, 7, 8 and 10

LEVERS FOR MANAGING THROUGHPUT

- ***Key concepts:*** Throughput, effective capacity, theoretical capacity, bottleneck resource
- ***Key management activity:*** Identify and manage bottleneck resource(s), throughput improvement mapping
- ***Key metric:*** Flow units per unit of time, Contribution margin per unit time

Levers for throughput management depend on one's location on the throughput improvement map:

1. If the throughput is significantly less than capacity, we say that the bottleneck is external. In that case the process is limited by factors that lie outside its bounds, such as the demand for its outputs or the supply of its inputs.
2. If the throughput is about equal to capacity, we say that the bottleneck is internal. In this case the only way to increase throughput is by increasing capacity. This can be done in two ways:
   a. Increase the financial capacity of the process, by modifying the product mix (give priority to products with higher profit margins)
   b. Increase the physical capacity of the process. This can be done in several ways depending on the situation:
      - If capacity is about equal to theoretical capacity, then existing resources are very efficiently utilized, and extra capacity

will require increasing the level of resources. The levers available in this case are:
- Increase the number of units
- Increase the size of resource units
- Increase the time of operation
- Subcontract or outsource
- Speed up the rate at which activities are performed

◦ If capacity is significantly lower than the theoretical capacity, then the existing resources are not utilized effectively, and the key to extra throughput is the elimination of capacity waste. The levers available in this case include:
- Eliminate non-value-adding activities
- Avoid defects, rework, and repetitions
- Reduce Time availability loss
- Reduce setup loss
- Move some of the work to non-bottleneck resources

LEVERS FOR REDUCING WAITING TIME

- ***Key concepts:*** Waiting time, buffer, variability, flow-time efficiency, cycle inventory, safety inventory, safety capacity
- ***Key management activity:*** Manage buffers to reduce waiting time
- ***Key metric:*** Waiting time in buffers

Buffers build up primarily because of batching or variability. The basic approaches to reducing waiting time can be summarized as follows:

1. Reduce cycle inventory (reduce batch size):
   - Reduce setup or order cost per batch
   - Renegotiate quantity discount policy
   - Reduce forward buying
2. Reduce safety inventory:
   - Reduce demand variability through improved forecasting
   - Reduce the replenishment lead time
   - Reduce the review period length
   - Reduce the variability in replenishment lead time
   - Pool safety inventory for multiple locations or products through either physical/virtual centralization or specialization or some combination there of
   - Exploit product substitution
   - Use common components
   - Postpone the product differentiation closer to the point of demand
3. Manage safety capacity:
   - Increase safety capacity
   - Decrease variability in arrivals and service
   - Pool available safety capacity
4. Synchronize flows:
   - Manage capacity to synchronize with demand
   - Manage demand to synchronize with available capacity
   - Synchronize flows within the process
5. Manage the psychological perceptions of the customers to reduce the cost of waiting

LEVERS FOR CONTROLLING PROCESS VARIABILITY

- ***Key concepts:*** Normal and abnormal variability, control limits, process capability, robust design
- ***Key management activity:*** Monitor and control process performance dynamically over time, reduce variability and its effects, design processes and products with low variability
- ***Key metrics:*** Quality, cost, time, inventory

1. Measure, prioritize, and analyze variability in key performance measures over time
2. Feedback control to limit abnormal variability:
   - Set control limits of acceptable variability in key performance measures
   - Monitor actual performance and correct any abnormal variability
3. Improve process capability:
   - Ensure that process performance is appropriately centered
   - Reduce normal variability
   - Design for producibility (simplify, standardize, and mistake-proof)
4. Immunize product performance to process variability through robust design

LEVERS FOR MANAGING FLOWS IN PROCESSING NETWORKS

- ***Key concepts:*** Waste, non-value-adding activities, product (cellular) layout, demand pull, quality at source, resource flexibility, employee involvement, supplier partnership, bullwhip effect, information flows, incentives, level production, river analogy, continuous improvement, reengineering

- ***Key management activity:*** Synchronize process flows while maintaining efficiency, set up a framework for process improvement
- ***Key metric:*** Cost, quality, time, flexibility

1. Managing flows in a plant:
- Process structure: Cellular layout
- Information and material flow: Demand pull system
- Level production: Batch size reduction
- Quality at source: Defect prevention, visibility, and decentralized control
- Resource flexibility: Reduce changeover times, cross train workforce
- Reduce processing variability: Standardization of work, process maintenance, maintenance of safety capacity
- Visibility of performance
- Supplier management: Partnership with information sharing and aligned incentives
- Human resource management: Employee involvement

2. Managing flows in a supply chain:
- Reduce information and material flow times through technology and efficient logistics
- Reduce fixed costs of ordering and quantity discounts
- Share information on customer demand and product availability
- Coordinate forecasts and replenishment decisions between various parties
- Stabilize prices

3. Improving processes:
- Frameworks: Continuous improvement and reengineering
- Tools: Increased visibility, incentives, improvement engine (PDCA cycle), and benchmarking

APPENDIX II

Probability Background

A. RANDOM VARIABLES

A random variable (*r.v.*) is a numerical outcome whose value depends on chance. We denote random variables with boldface uppercase letters, and their realized values with lowercase italicized letters. For example, suppose demand for sweaters next winter cannot be predicted with certainty, and we think it could be anywhere between 0 and 1000. Then we may denote the uncertain demand by *r.v.* **X**, which takes on values in the set $\{0, 1, \cdots, 1000\}$. If at the end of the season, the actual demand turns out to be 734 units, we say the value of **X** is $x = 734$.

1. Probability Distribution

The expression $\{\mathbf{X} \leq x\}$ is the uncertain event that the *r.v.* **X** takes on a value less than or equal to x. The event is uncertain because whether it occurs or not depends on the value of **X**. The probability that the event occurs is denoted as $\Pr\{\mathbf{X} \leq x\}$. As x varies, this probability defines a function:

$$F(x) = \Pr\{\mathbf{X} \leq x\}, \qquad -\infty < x < \infty,$$

which is called the "cumulative distribution function" (c.d.f.) of the *r.v.* **X**. Sometimes, we write it as $F_{\mathbf{X}}(x)$ to highlight that it is distribution function of **X**. The cumulative distribution of a random variable contains all information about it.

An *r.v.* **X** is called "discrete" if it can take on only a finite or countable number of values $x_1, x_2, \ldots$ with probabilities $p_i = \Pr\{\mathbf{X} = x_i\}$ for $i = 1, 2, \ldots$ and $\Sigma_i p_i = 1$. The function

$$f(x_i) = \Pr\{\mathbf{X} = x_i\} = p_i, \qquad \text{for } i = 1, 2, \ldots,$$

is called the "probability mass function" of the discrete random variable **X**. It is related to the cumulative distribution function as

$$F(x) = \sum_{x_i \leq x} f(x_i).$$

A random variable **X** is called "continuous" if it takes on a continuum of values x. In that case, it is improbable that it will take on any specific value x, i.e., $\Pr\{\mathbf{X} = x\} = 0$ for every x. Then its cumulative distribution function is continuous in x. Often there exists a probability density function $f(x)$ such that

$$F(x) = \int_{-\infty}^{x} f(u)du$$

which is the area under the probability density function to the left of x.

2. Expected Value or Mean

The expected value or the mean of a random variable **X** is the weighted average of all of its possible values, using probabilities as weights

$$E(\mathbf{X}) = \begin{cases} \sum_i x_i f(x_i) & \text{if } \mathbf{X} \text{ is a discrete r.v.} \\ \int_{-\infty}^{\infty} uf(u)du & \text{if } \mathbf{X} \text{ is a continuous r.v.} \end{cases}$$

We will also denote the mean of **X** by the italicized font X or by μ_X.

3. Variance and Standard Deviation

The variance of a random variable **X** is a measure of its variability from the mean. It is computed as the expected squared deviation of **X** from its mean μ_x and is denoted by

$$V(\mathbf{X}) = E[(\mathbf{X} - \mu_{\mathbf{x}})^2].$$

Standard Deviation

The square-root of the variance of a *r.v.* **X** is called its "standard deviation" and is denoted by

$$\sigma_{\mathrm{X}} = \sqrt{V(\mathbf{X})} = \sqrt{E[(\mathbf{X} - \mu_{\mathbf{x}})^2]}.$$

Coefficient of Variation

The coefficient of variation of a random variable **X** measures its standard deviation relative to its mean and is denoted by

$$C_X = \frac{\sigma_X}{\mu_X}.$$

The variance, standard deviation, and the coefficient of variation are all measures of the amount of uncertainty or variability in **X**.

With these basic definitions relating to single random variables, we can now consider some concepts and results that involve multiple random variables.

4. Independence

Two random variables $\mathbf{X}_1$ and $\mathbf{X}_2$ are said to be (mutually) independent if knowing the value of one does not change the probability distribution of the other. Formally, two random variables $\mathbf{X}_1$ and $\mathbf{X}_2$ are said to be independent if and only if for any two events A and B,

$$\Pr(\mathbf{X}_1 \in A \text{ and } \mathbf{X}_2 \in B) = \Pr(\mathbf{X}_1 \in A)\Pr(\mathbf{X}_2 \in B).$$

Otherwise, $\mathbf{X}_1$ and $\mathbf{X}_2$ are said to be dependent. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent then it follows that

$$E(\mathbf{X}_1\mathbf{X}_2) = E(\mathbf{X}_1)E(\mathbf{X}_2).$$

5. Covariance and Correlation Coefficient

Suppose $\mathbf{X}_1$ and $\mathbf{X}_2$ are two random variables with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, respectively. The covariance of $\mathbf{X}_1$ and $\mathbf{X}_2$ is defined as the expected value of the product of their deviations from their respective means and is denoted by

$$Cov(\mathbf{X}_1, \mathbf{X}_2) = E[(\mathbf{X}_1 - \mu_1)(\mathbf{X}_2 - \mu_2)].$$

The correlation coefficient is then defined as

$$\rho = \frac{Cov(\mathbf{X}_1, \mathbf{X}_2)}{\sigma_1\sigma_2}.$$

The value of the correlation coefficient is always between −1 and +1. A positive covariance or correlation coefficient implies that the two *r.v.*s tend to vary in the same direction (up or down). Similarly, negative covariance or correlation coefficient implies that on average they tend to move in the opposite direction. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent then the two are uncorrelated, or

$$Cov(\mathbf{X}_1, \mathbf{X}_2) = 0.$$

6. Sums of Random Variables

Consider two random variables $\mathbf{X}_1$ and $\mathbf{X}_2$. Then, it turns out that

$$E(\mathbf{X}_1 + \mathbf{X}_2) = E(\mathbf{X}_1) + E(\mathbf{X}_2)$$
$$V(\mathbf{X}_1 + \mathbf{X}_2) = V(\mathbf{X}_1) + V(\mathbf{X}_2) + 2Cov(\mathbf{X}_1, \mathbf{X}_2)$$

Recall that if $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent, then $Cov(\mathbf{X}_1, \mathbf{X}_2) = 0$. It then follows that the expected value and variance of sums of independent random variables is equal to the sum of their expectations and variances, respectively.

If $\mathbf{X}_1$ and $\mathbf{X}_2$ have identical distributions (and therefore same mean μ and standard deviation σ), with a correlation coefficient ρ, then the standard deviation of the sum $\mathbf{X}_1 + \mathbf{X}_2$ is

$$\sigma_{\mathbf{X}_1+\mathbf{X}_2} = \sqrt{2(1+\rho)}\sigma.$$

B. PROBABILITY DISTRIBUTIONS

We consider specific discrete and continuous probability distributions that arise in operations management and other applications.

1. Binomial Distribution

Suppose the uncertain outcome of a trial (such as a project) is either success, with probability p, or failure with probability $(1 - p)$. We may define a binary *r.v.* $\mathbf{N_1}$, with $\mathbf{N_1} = 1$ corresponding to success and $\mathbf{N_1} = 0$ denoting failure. Then we can compute its mean $\mathrm{E}(\mathbf{N_1}) = p$ and variance $V(\mathbf{N_1}) = p(1 - p)$.

Now suppose we conduct n independent identically distributed (*i.i.d.*) trials, each with probability of success p and count the total number of successes $\mathbf{N = N_1 + N_2 + \ldots + N_n}$. Then the discrete r.v. $\mathbf{N}$ is said to have a binomial distribution with parameters n and p. Its mean is given by $E(\mathbf{N}) = np$ and variance is $V(\mathbf{N}) = np(1 - p)$, since the expected value of the sum is equal to the sum of the expected values and the same is true of the variance due to independence. Probability mass function of the binomial distribution is given by

$$\Pr\{\mathbf{N} = k\} = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0, 1, 2, \ldots, n$$

To see the logic, note that p^k is the probability of observing k successes in n independent trials, $(1 - p)^{n-k}$ is the probability of observing remaining $(n - k)$ failures, and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the number of combinations of observing k successes and $(n - k)$ failures. This probability can be computed in Microsoft Excel using BINOMDIST function as

$$\Pr\{\mathbf{N} = \mathrm{k}\} = \text{BINOMDIST}(k,n,p,\text{False}),$$

and the cumulative probability can be computed as

$$\Pr\{\mathbf{N} \le \mathrm{k}\} = \text{BINOMDIST}(k,n,p,\text{True}),$$

where "True" and "False" can also be expressed by 1 and 0, respectively.

2. Poisson Distribution

It turns out that if n is "large", and p is "small", the binomial distribution can be approximated by the Poisson distribution. It serves as a useful model in a wide variety of operations applications, such as number of calls received at a call center, number of defects produced by a process, number of customer complaints received, etc. In general, consider a sequence of events that occur randomly through time. Suppose the average rate of occurrence is R events per unit of time. Let $\mathbf{N}(t)$ be the discrete random variable representing the number of events that occur in a time interval of duration t. We say that $\mathbf{N}(t)$ has a Poisson distribution with mean Rt, if

$$\Pr\{\mathbf{N}(t) = n\} = e^{-Rt}\frac{(Rt)^n}{n!}, \text{ where } n = 0, 1, 2, \ldots$$

The mean number of events in time period t is given by $E[\mathbf{N}(t)] = Rt$, which also happens to be the variance of the number of events in time period t. The Poisson probabilities can be calculated in Microsoft Excel with the POISSON function as follows:

$$\Pr\{\mathbf{N}(t) = n\} = \text{POISSON}(n, Rt, \text{False})$$

$$\Pr\{\mathbf{N}(t) \le n\} = \text{POISSON}(n, Rt, \text{True}).$$

3. Exponential Distribution

Suppose random events (such as customer arrivals) occur over time according to the Poisson processs at rate R. Then it turns out that the time $\mathbf{T}$ between two consecutive events will have exponential distribution with mean $m = 1/R$, and its probability density function is given by

$$f(t) = Re^{-Rt}, t \ge 0$$

The mean and the standard deviation of the elapsed time between consecutive events is also m, so the coefficient of variation of an exponential random variable is equal to 1. The exponential probability density function and cumulative distributions can also be evaluated in Microsoft Excel with the EXPONDIST function as follows:

$$f(t) = \text{EXPONDIST}(t, R, \text{False}),$$

$$F(t) = \text{EXPONDIST}(t, R, \text{True}).$$

4. Normal Distribution

This is perhaps the most important distribution in probability and statistics. It often arises in practice due to a result in probability known as "The Central Limit Theorem", which roughly states that sums and averages of independent identically distributed random variables tend to be normally distributed. For example, if n is "large" and p is "not small", then binomial (n, p) distribution can be approximated well by the normal distribution with mean np and variance $np(1 - p)$. A normal distribution is completely characterized by its mean μ and standard deviation σ and has the probability density function given by:

$$f_{\mathbf{X}}(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Normal random variable is often denoted as $N(\mu, \sigma)$. The probability density function of normal distribution is bell shaped and symmetric around its mean.

In applications, one is often interested in computing the probability that a normal random variable is smaller than a given value and vice versa. We indicate two computational methods: one uses Microsoft Excel functions, while the other is based on the traditional approach of standardizing a normal random variable. Given a *r.v.* $\mathbf{X}$ that is $N(\mu, \sigma)$, we often face two problems.

PROBLEM 1. Given any number x, find a probability $p = \Pr\{\mathbf{X} \le x\}$.

Method 1. Use the Microsoft Excel function NORMDIST:

$$p = \text{NORMDIST}(x, \mu, \sigma, \text{True}).$$

Method 2. Transform the random variable $\mathbf{X}$ and the number x into a new random variable $\mathbf{Z}$ and number z as:

$$\mathbf{Z} = \frac{\mathbf{X} - \mu}{\sigma} \quad \text{and} \quad z = \frac{x - \mu}{\sigma}$$

Then $\mathbf{Z}$ is also normally distributed with mean 0 and a standard deviation 1, i.e., $\mathbf{Z} = N(0, 1)$, which is called the "standard normal" random variable. Clearly

$$p = \Pr\{\mathbf{X} \le x\} = \Pr\{\mathbf{Z} \le z\},$$

which can be read from the table of cumulative distribution of the standard normal *r.v.* given in Table A.1. For negative values of z, notice by symmetry that $\Pr(\mathbf{Z} \le z) = 1 - \Pr(\mathbf{Z} \le -z)$.

PROBLEM 2. Conversely, given any probability p, find a number x, such that $\Pr\{\mathbf{X} \le x\} = p$.

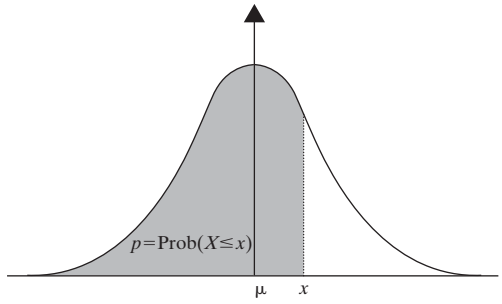


FIGURE A.1 A Graphic Representation of the Probability p (Shaded Area) that a Normal Random Variable X is Less Than or Equal to a Number x.

Method 1. Use the Microsoft Excel function NORMINV:

$$x = \text{NORMINV}(p, \mu, \sigma).$$

Method 2. Transform the random variable **X** and the number x into their "standard" counterparts **Z** and z as described above. Given p, read z backwards from Table A.1 such that $\Pr(\mathbf{Z} \leq z) = p$ and find the quantity x by transforming back:

$$x = \mu + z\sigma.$$

Table A.1 The Cumulative Standard Normal Distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997

This table represents the cumulative distribution function of a standard normal random variable. That is, it gives the probability p that the standard normal random variable $\mathbf{N}(0,1)$ is less than or equal to a quantity z. For example, if $z = 1.65$, then $p = 0.9505$.

SOLUTIONS TO SELECTED PROBLEMS

Chapter 3: Process flow measures

Exercise 3.1 (Bank)

For the bank we have

Average inventory $I = 10$ people,

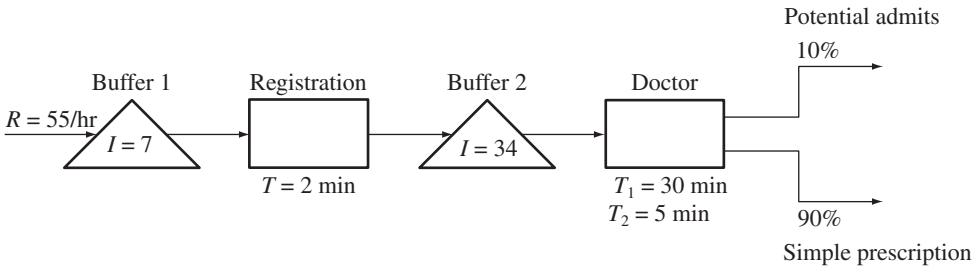
Throughput $R = 2$ people/min (we assume a stable system).

Thus,

Average wait time $T = I / R = 10/2$ min $= 5$ min.

Exercise 3.4 (ER)

First, draw the flowchart with all the data given:



We assume a stable system. This implies that average inflow equals average outflow at every stage. In this case you are given inventory numbers I and flow rate $R = 55$ patients/hr. There are two flow units:

(1) Those that are potential admits: flow rate $= 55 \times 10\% = 5.5$/hr.

(2) Those that get a simple prescription: flow rate $= 55 \times 90\% = 49.5$/hr.

To find the average flow times, we use Little's law at each activity for which the flow time is unknown:

(1) ***Buffer 1:*** $R = 55$/hr (both flow units go through there), $I = 7$, so that waiting time in buffer 1 $= T = I/R = 7/55$ hr $= 0.127$ hour $= 7.6$ minutes.

(2) ***Registration:*** flow time $T = 2$ min $= 2/60$ hr. All flow units flow through this stage. Thus, flow rate through this stage is $R = 55$/hr. Average inventory at registration is given by $I = RT = 55 \times 2/60 = 1.83$ patients.

(3) ***Buffer 2:*** $R = 55$/hr (both flow units go through there), $I = 34$, so that waiting time in buffer 2 $= T = I/R = 34/55$ hr $= 0.62$ hour $= 37.1$ minutes.

(4) ***Doctor time:*** depends on the flow unit:

4a: potential admits: $T = 30$ minutes

4b: prescription folks: $T = 5$ minutes

OK, now we have everything to find the total average flow times: find the critical path for each flow unit. In this case, each flow unit only has one path, so that is the critical path. We find its flow time by adding the activity times on the path:

a. For a potential admit, average flow time (buffer 1 + registration + buffer 2 + doctor) $= 7.6 + 2 + 37.1 + 30 = 76.7$ minutes

b. For a person ending up with a prescription, average flow time (buffer 1 + registration + buffer 2 + doctor) $= 7.6 + 2 + 37.1 + 5 = 51.7$ minutes.

The answer to the questions is found as follows:

(1) *On average, how long does a patient spend in the emergency room?*

We know the flow time of each flow unit. The average flow time over all flow units is the weighted average: 10% of total flow units spend 76.7 minutes while 90% spend 51.7 minutes. Thus, the grand average is:

$$T = 10\% \times 76.7 + 90\% \times 51.7 = 54.2 \text{ minutes.}$$

(2) *On average, how many patients are being examined by a doctor?*

This question asks for the average inventory at the doctor's activity. Again, first calculate inventory of each type of flow unit:

a. Potential admits: $R = 5.5$ patients/hr, $T = 30$ min $= 0.5$ hr, thus, $I = RT = 5.5$/hr $\times 0.5$ hr $= 2.75$ patients

b. Simple prescription: $R = 49.5$ patients/hr, $T = 5$ min $= (5/60)$ hr, thus $I = RT = 49.5 \times (5/60) = 4.125$ patients

Thus, total inventory at the doctor is $2.75 + 4.125 = 6.875$ patients.

(3) *On average, how many patients are in the ER?*

This question asks for total inventory in ER = inventory in buffer 1 + inventory in registration + inventory in buffer 2 + inventory with doctors $= 7 + 1.83 + 34 + 6.875 = 49.705$ patients.

Chapter 4: Flow Time Analysis

Exercise 4.1

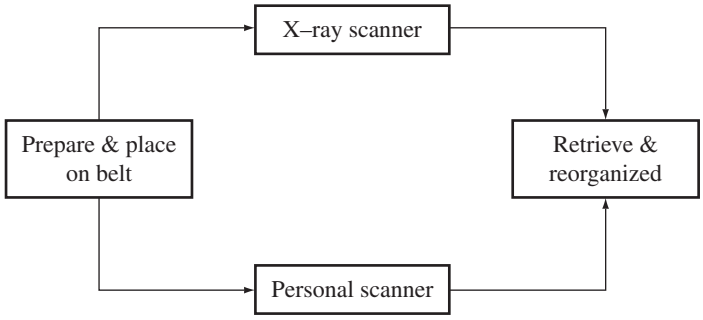
a. The first step is to find the average flow time for the activities "time with judge" and "pay fine." This can be done by averaging the ten observations given in the table. These amount to 1.55 minutes and 2.4 minutes, respectively. Thus, the theoretical flow time is 1.55 + 2.4 = 3.95 minutes.

b. The average flow time of the process can be estimated by taking the average of the total time each defendant spent in the system. This comes out to 120 minutes.

c. The flow time efficiency is 3.95/120 = 3.3%

Exercise 4.5

a.



b. The critical path is the one through the X-ray scanner.

The flow time is 30 + 40 × 1.5 + 60 = 150 seconds.

c. The flow time efficiency is 150/530 = 28%

d. The personal scanner is not on the critical path.

Chapter 5: Flow Rate and Capacity Analysis

Exercise 5.2

a. The capacity of shopping cases is 24/4 = 6 cases per day and the capacity medical cases is 24/6 = 4 cases per day.

The contribution from shopping cases is 6 × 4000 = $24,000 per day and from medical cases is 4 × 5000 = $20,000 per day. Shopping is more profitable.

b. The capacity is 4.8 cases per day, and the margin per unit is 0.5 × 5000 + 0.5 × 4000 = $4500 per case. The margin at capacity is 4.8 × 4500 = $21,600 per day

c. The profit at capacity is (21,600 × 20) − 500,000 = −68,000 (a loss)

d. An extra paralegal will increase the capacity by 120/5= 24 cases per month. This is worth 24 × 4500 = $108,000 per month (less the cost of the paralegal)

Exercise 5.3

a. The capacity is 6 customers per hour:

The unit load from the three hair stylists is 10 + 15 + 5 = 30 minutes per customer. Checking in (LuLu) takes 3 minutes. Thus, the capacity of the stylists is 3 × 60/30 = 6 customers per hour, and of LuLu is 1 × 60/3 = 20 customers per hour. The bottleneck are the stylists.

Exercise 5.6

a. The variable cost is 33% × 15 = $5 million per month. The throughput profit multiplier is (18−5)/(18−15) = 13/3 = 4.3

Chapter 6: Inventory Analysis

Exercise 6.2

BIM Computers: Assume 8 working hours per day.

a. We know Q = 4 wks supply = 1,600 units; R = 400 units/wk = 20,000 units/yr; purchase cost per unit C = 80% × $1250 = $1,000. Thus, holding cost $H = rC$ = 20%/year × $1,000 = $200/yr. Switch over or setup cost S = $2,000 + (1/2 hr × $1,500/day × 1 day/8 hr) = $2,093.75. Thus, number of setups per year = R/Q = 20,000 units/yr/1600 units/setup = 12.5 setups/yr. Thus,

- Annual setup cost = $(R/Q) \times S$ = 12.5 setups/yr × $2,093.75/setup = $26,172/yr.
- Annual Purchasing Cost = $R \times C$ = 20,000 units/yr × $1,000/unit = $ 20 M/yr.
- Annual Holding Cost = $(Q/2) \times H$ = 800 × $200/yr = $160,000/yr.
- Thus, total annual production and inventory cost = $20,186,172.

b. The economic order quantity,

$$\text{EOQ} = \sqrt{\frac{2RS}{H}} = \sqrt{\frac{2 \times 20000 \times 2093.75}{200}} = 647 \text{ units.}$$

The associated calculations are as follows:

- Number of setups = R/Q = 20,000/647 = 30.91. Thus, annual setup cost = 30.91 setups/yr × $2,093.75/setup = $64,718/yr.
- Annual holding cost = $(Q/2) \times H$ = 323.6 × $200/yr = $64,720/yr
- Annual purchasing cost remains $20M/yr
- The resulting annual savings equals $20,186,172 − $20,129,438 = $56,734.

Exercise 6.4

Current fixed costs, say, S_1 = \$1000. Current optimal lot size Q_1 = 400. New, desired lot size Q_2 = 50. Intuitively, since the lot size needs to decrease by a factor of 8 and demand is unchanged, the fixed costs need to go down by a factor of $1/(8)^2 = 1/64$. Thus, the new fixed cost should be \$1000/64 = \$15.625. Formally, we must find the fixed cost S_2 at which Q_2 is optimal. Since Q_1 is optimal for S_1, we have

$$Q_1 = 400 = \sqrt{\frac{2RS_1}{H}} = \sqrt{\frac{2 \times R \times 1000}{H}}$$

So, $R/H = 160000/2000 = 80$. Now,

$$Q_2 = 50 = \sqrt{\frac{2RS_2}{H}}$$

or $S_2 = 50^2/(2 \times 80) = 15.625$. So, the retailer should try to reduce her fixed costs to \$15.625.

Exercise 6.10

Each retail outlet faces an annual demand, $R = 4000/\text{wk} \times 50 = 200{,}000$ per year. The unit cost of the item, C = \$200/unit. The fixed order cost, S = \$900. The unit holding cost per year, $H = 20\% \times 200$ = \$40/unit/year.

a. The optimal order quantity for each outlet

$$Q = \sqrt{\frac{2RS}{H}} = \sqrt{\frac{2 \times 200{,}000 \times 900}{40}} = 3000$$

with a cycle inventory of 1500 units. The total cycle inventory across all four outlets equals 6000 units.

b. With centralization of purchasing the fixed order cost, S = \$1800. The centralized order quantity is then,

$$Q = \sqrt{\frac{2RS}{H}} = \sqrt{\frac{2 \times 800{,}000 \times 1800}{40}} = 8485$$

and a cycle inventory of 4242.5 units.

Chapter 7: Managing Flow Variability: Safety Inventory

Exercise 7.2

a. Average weekly demand (R) = 1000
Standard deviation of weekly demand (σ_R) = 150.
Lead time (L) = 4 weeks.
Standard deviation of demand during lead time $\sigma_{LTD} = \sqrt{L}\sigma_R = 300$.
Current reorder point (ROP) = 4,200.
Average demand during lead time (LTD) = $L \times R = 4{,}000$.
Current level of safety stock (I_{safety}) = 200.
Current order quantity (Q) = 20,000
Average inventory (I) = $I_{safety} + Q/2 = 200 + (20{,}000/2) = 10{,}200$.
Average time in store (T) = $I/R = 10{,}200/1{,}000 = 10.2$ weeks.
To estimate total costs, observe that the fixed cost per order, S = \$100 and the holding cost per unit per year. H = 25%/year times \$1/year = \$0.25/year
Annual ordering cost = $S \times R/Q$ = \$100 × 2.5 = \$250.
Annual holding cost = $H \times I$ = \$0.25 × 10,200 = \$2,550.

b. We use the EOQ formula to determine the optimal order quantity.

$$Q = \sqrt{\frac{2RS}{H}} = \sqrt{\frac{2 \times 50{,}000 \times 100}{0.25}} = 6{,}325$$

To determine the safety inventory, I_{safety}, for a 95% level of service, we first observe that the z-value = 1.65. Then $I_{safety} = z \times \sigma_{LTD} = 1.65 \times 300 = 495$.

Then, average inventory (I) = $I_{safety} + Q/2 = 495 + (6{,}325/2) = 3{,}657.5$.
Average time in store (T)= $I/R = 3.6575$ weeks.

c. If lead time (L) reduces to 1 week, then standard deviation of demand during lead time (σ_{LTD}) = 150.
Safety stock for 95% level of service = 1.65 × 150 = 247.5.
Average inventory = 247.5 + (6,325/2) = 3,410.
Average time in store = 3.41 weeks.

Exercise 7.5

The revenue per crate, p = \$120.00; variable cost, c = \$18.00; and salvage value, $v = -\$2.00$. The marginal benefit of stocking an additional crate (MB) = $p - c$ = \$120 − \$18 = \$102. The marginal cost of stocking an additional unit (MC) = $c - v$ = \$18 + \$2 = \$20. Then,

$$\text{MB}/(\text{MB} + \text{MC}) = 102/(102 + 20) = 0.836.$$

The probability density of demand and its cumulative probability is listed below.

Demand	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frequency	0	0	0	1	3	2	5	1	6	7	6	8	5	4	1	3
Prob.	0	0	0	0.02	0.06	0.04	0.1	0.02	0.12	0.13	0.12	0.15	0.1	0.08	0.02	0.06
Cumulative Prob.	0	0	0	0.02	0.08	0.12	0.21	0.23	0.35	0.48	0.6	0.75	**0.85**	0.92	0.94	1

The optimal order quantity is the smallest number of crates such that cumulative probability is at least 0.836. From the table this gives the number of crates to be 12.

Exercise 7.8

a. To compute the optimal order quantity at each store we use the EOQ formula.
Assume 50 sales weeks/year; H = 25%/year times \$10.
R = 10,000/week = 500,000/year and S = \$1000. Thus,

$$(optimal)\ Q = \sqrt{\frac{2RS}{H}} = \sqrt{\frac{2 \times 500{,}000 \times 1000}{2.5}} = 20{,}000$$

The replenishment lead time (L) = 1 week.
Standard deviation of demand during lead time at each store (σ_{LTD}) = 2,000.
Safety stock at each store for 95% level of service (I_{safety}) = 1.65 × 2,000 = 3,300.
Reorder point (ROP)= $R \times L + I_{safety}$ = 10,000 + 3,300 = 13,300.
Average inventory across four stores (I^d) = 4 × $(I_{safety} + Q/2) = 4 \times (3{,}300 + (20{,}000/2)) = 53{,}200$.

Annual order cost for all four stores = $4 \times S \times R/Q$ = 4 × 1,000 × 25 = \$100,000.
Annual holding cost for all four stores = $H \times I^d$ = \$133,000.
Average time unit spends in store (T) = I^d / 4 × R = 53,200/40,000 = 1.33 weeks.

b. To compute the optimal order quantity at centralized store observe that this store faces a cumulative average weekly demand = 4 × 10,000 = 40,000. This gives an annual demand of 2,000,000 units.

$$(optimal)\ Q = \sqrt{\frac{2RS}{H}} = \sqrt{\frac{2 \times 2{,}000{,}000 \times 1000}{2.5}} = 40{,}000$$

Standard deviation of demand during lead time at central store

$$\sigma_{LTD} = \sqrt{4} \times 2000 = 4000$$

Safety stock at central store for 95% level of service = 1.65 × 4,000 = 6,600.
Reorder point (ROP) = $R \times L + I_{safety}$ = 40,000 + 6,600 = 46,600.
Average inventory in central store (I^c) = 6,600 + (40,000/2) = 26,600.
Annual order cost for central store = $S \times R/Q$ = \$1,000 × 50 = \$50,000.
Annual holding cost for central store = $H \times I^c$ = \$66,500.
Average time unit spends in the central store = I^c/R = 26,600/40,000 = 0.67.

Exercise 7.10

Mean demand is, 1000/day with a daily standard deviation 150.
Annual unit holding cost, H = 0.25 × \$2/unit/year = \$5/unit/year.
Review period, T_r = 2 weeks and replenishment lead time, L = 1 week.

a. Average weekly demand, R = 7×1000 = 7,000; weekly standard deviation of demand = $\sigma_R = \sqrt{7} \times 150 = 397$
Standard deviation of demand during review period and replenishment lead time

$$\sigma_{RLTD} = (\sqrt{T_r + L})\sigma_R = \sqrt{3} \times 397 = 688$$

For a 98% service level z = NORMSINV (0.98) = 2.054 and safety stock

$$I_{safety} = z \times \sigma_{RLTD} = 2.054 \times 688 = 1413$$

$$\text{OUL} = R \times (T_r + L) + I_{safety} = 7000 \times 3 + 1413 = 22{,}413 \text{ units}$$

Average order quantity, $Q = R \times T_r$ = 14,000 and therefore cycle stock = 14,000/2 = 7,000.
Average inventory, $I = Q/2 + I_{safety}$ = 7,000 + 1,413 = 8,413.
Total average annual holding cost = $H \times I$ = 5.0 × 8,413 = \$42,065 per unit per year.

b. If review period, T_r, is reduced 1 week, then, standard deviation of demand during review period and replenishment lead time

$$\sigma_{RLTD} = (\sqrt{T_r + L})\sigma_R = \sqrt{2} \times 397 = 562$$

For a 98% service level z = NORMSINV (0.98) = 2.054 and safety stock

$$I_{safety} = z \times \sigma_{RLTD} = 2.054 \times 562 = 1154$$

$$\text{OUL} = R \times (T_r + L) + I_{safety} = 7000 \times 2 + 1154 = 15{,}154 \text{ units}$$

Average order quantity, $Q = R \times T_r$ = 7,000 and therefore cycle stock = 7,000/2 = 3,500.
Average inventory, $I = Q/2 + I_{safety}$ = 3,500 + 1,154 = 4,654.
Total average annual holding cost = $H \times I$ = 5.0 × 4,654 = \$23,270 per unit per year.
Total savings in holding costs = \$42,065 − \$23,270 = \$18,795 per year.
Of course, now we order twice as frequently. So any associated costs related to placing orders needs to be balanced off against the savings in inventory holding costs of a shorter review period.

Chapter 8: Managing Flow Variability: Safety Capacity

Exercise 8.5 (Heavenly Mercy Hospital)

a. We are given:

Average arrival rate, R_i = 18 per hour = 0.3 per minute
Average unit capacity, $1/T_p$ = 2 per hour = 1/30 per minute

Cost per server = \$100 per hour

Desired average time in system = 40 minutes.

To determine the number of servers c, we know that capacity utilization should be less than 100%, so

Utilization = inflow/capacity = $18/2c < 1$, or $c > 9$

With infinite buffer capacity, the Performance spreadsheet provides the following results:

Number of Servers (c)	Average Number in the System (I)	Average Time in System (T)
10	15	50 minutes
11	11	36.5 minutes
12	10	32.7 minutes

Thus hiring 11 servers limits the average turnaround time of 36.46 minutes which is under the desired target of 40 minutes, at the hourly cost of \$1,100.

Cost of customer waiting =

\$1 per minute = \$60/hour = \$60/hour

Cost per server = \$20 per hour.

To determine the number of servers c, we know that we should have a utilization of less than 100%, or

Utilization = inflow/capacity =
$52/\text{hr}/(c \times 20/\text{hr}) < 1$ so that $c > 2.6$.

Increasing the number of servers from 3 upward, the Performance spreadsheet with infinite buffer capacity yields:

Number of Servers c	Server cost per hour	Average Queue length I_i	Average waiting time T_i	Waiting cost per hour = \$60I_i	Total cost per hour
3	\$60	4.94	5.69	\$296.1	\$356.10
4	\$80	0.66	0.76	\$39.5	\$119.6
5	\$100	0.16	0.19	\$9.6	\$109.6
6	\$120	0.043	0.049	\$2.56	\$122.56

b. Now suppose the service time is reduced to 20 minutes but the equipment and radiologist costs \$150 per hour. Then we have

Average arrival rate, $R_i = 18$ per hour = 0.3 per minute,

Average unit capacity, $1/T_p = 3$ per hour = 1/20 per minute,

Cost per server = \$150 per hour,

Desired average time in system = 40 minutes.

To determine the number of servers c, we should ensure the utilization of less than 100%, so

Utilization = inflow/capacity = $18/3c < 1$, or $c > 6$

With $c = 7$, the average time in the system is $T = 32.27$ minutes, which is less than 40 minutes desired.

The cost of hiring seven servers at \$150 per hour is = 7×150 = \$1,050. Thus, it is advantageous to lease the more sophisticated equipment. It reduces the total cost as well as the overall time in the system.

Exercise 8.7 (Global Airlines)

We are given:

Average arrival rate, $R_i = 52$ per hour = 52/60 per minute,

Average unit capacity, $1/T_p = 20$ per hour = 1/3 per minute,

Thus Global should staff with 5 agents. Observe that the industry norm of averaging under 3 minutes of waiting can be achieved using only 4 agents.

Exercise 8.9 (Burrito King)

a. Reducing variability will reduce the waiting time

b. Server utilization will remain the same, since the average demand and service rates do not change.

Exercise 8.11 (Master Karr)

a. Waiting cost = (\$60/hr) × (average # of customers waiting) = \$60I_i = \$94.65/hr

Blocking cost = (\$50/call) × (average # of busy calls/hr) = \$50/call × $R \times P_b$ = \$149.93/hr

Staffing cost = \$15/CSR × (# of CSRs) = \$15c = \$75/hr

Total cost = \$319.58/hr

b. Now the number of servers c increases to 6 while the buffer capacity K decreases to 9. Recalculate the performance with the spreadsheet (calculations summarized above).

Waiting cost = \$60I_i = \$30.77/hr

Blocking cost = \$50/call × $R \times P_b$ = \$29.78/hr

Staffing cost = \$15c = \$90/hr

Total = \$150.54/hr

Part	Arrival Rate R_i	Service Time T_p	Number of Servers c	Buffer Capacity K	Capacity Utilization u	Probability of blocking P_b	Average queue length I_l
a	4	1	5	10	79.00%	1.25E-02	1.58
b	4	1	6	9	66.50%	2.48E-03	0.51

Hence, we should add a server, as it yields a cost savings of \$319.58 − \$150.54 = \$169.04/hr.

Exercise 8.13 (McBerger)

a. The average waiting time in queue will decrease, because less variability leads to less waiting, by the queue length formula.

b. The average waiting time in queue will decrease because shorter service time leads to lower capacity utilization and hence less waiting.

Chapter 9: Managing Flow Variability: Process Control and Capability

Exercise 9.1 (Costello Labs)

a. Given the symmetric shape of normal distribution around its mean, maximum conformance of the output within the given specifications will be achieved by centering the process at the midpoint of the specifications, that is, at $\mu = 32.5$ gms. Now if we desire 98% of the output to conform to the specifications, from the normal distribution tables, the specification limits should be $z = 2.33$ standard deviations on either side of the mean. Therefore, $(35 - 32.5)/\sigma = 2.33$, or $\sigma = 2.5/2.33 = 1.073$ gms. The corresponding process capability ratio is $C_p = (35 - 30)/6\sigma = 0.78$.

b. With sample size $n = 12$, and process precision $\sigma = 1.073$ as above, we can now determine the ideal control limits on subgroup averages of 12 bottles as:

$\overline{X}$ control chart: $\mu \pm 3\sigma/\sqrt{n} = 32.5 \pm (3)(1.073)/\sqrt{12} = (31.57, 33.34)$

Exercise 9.3 (First Chicago Bank)

From the 26 observations given, we can calculate the average number of errors per thousand transactions $m = 3.3077$, or the average fraction defective $p = 0.0033$, which is much better than the industry average of 0.015.

a. Note that the number of errors N in 1000 transactions has binomial distribution with $n = 1000$ and $p = 0.0033$. Since n is large and p is small, we can use Poisson approximation with mean $c = 3.3077$. Control limits on the number of errors can be determined as $c \pm 3\sqrt{c} = (0, 8.76)$.

b. Observe that three observations 4, 15, and 17 out of the 26 given exceed the UCL = 8.75. Hence, the process is *not* in control, even though on average the bank's performance is better than the BAI standard! The process is not stable, and our estimate of m is not reliable. We first need to stabilize the process by removing assignable causes.

Exercise 9.5

Given the process mean $\mu = 6$ cm, standard deviation $\sigma = 0.01$ cm, and sample size $n = 10$, the $\overline{X}$ chart control limits are $LCL = \mu - 3\sigma/\sqrt{10} = 5.9905$ and $UCL = \mu + 3\sigma/\sqrt{10} = 6.0095$ cm

Note that the customer specifications are irrelevant in setting the control limits.

Exercise 9.8 (Balding Inc.)

If D is the diameter of basketballs produced, we need Prob $(29.3 \leq D \leq 29.7) = 0.98$.

Now we know that Prob $(-2.33 \leq Z \leq 2.33) = 0.98$, so $z = (29.7 - 29.5)/\sigma = 2.33$ or we need $\sigma = 0.086$ inches.

Hence, the process capability ratio should be $C_p = (USL - LSL)/6\sigma = (29.7 - 29.3)/6\sigma = 0.7752$.

GLOSSARY

80-20 Pareto principle A principle that states that roughly 20% of problem types account for 80% of all occurrences.

Abandonment A situation when a customer, having waited in queue for some time, leaves the process before being served.

Abnormal variability Unpredictable variability that disturbs the state of statistical equilibrium of the process by changing parameters of its distribution in an unexpected way.

Activity The simplest form of transformation; the building block of a process.

Activity time The time required by a typical flow unit to complete the activity once. Also called the flow time of the activity.

Aggregation, principle of A statistical principle that states that the standard deviation of the sum of random variables is less than the sum of the individual standard deviations.

All-unit quantity discount policy A quantity discount policy where a buyer receives discount on all units purchased whenever the quantity purchased exceeds a certain threshold; also see *incremental unit discount policy*.

American system of manufacturing The manufacturing system begun in 1810 that introduced the use of interchangeable parts, thereby eliminating the need to custom fit parts during assembly.

Andon Literally, andon means a display board. In the Toyota Production System, a worker is empowered to stop the line by pulling a cord. A display board identifies the station that pulled the cord, enabling the supervisor to locate it easily.

Availability loss factor Resource availability loss as a fraction of scheduled availability.

Average flow rate The average number of flow units that flow through (into and out of) the process per unit of time. Also called *throughput.*

Average flow time The average of the flow times across all flow units that exit the process during a specific span of time.

Backlogged The situation in which customers must wait to have their demand satisfied.

Batch The size of an order or production in response to the economies of scale.

Benchmarking The process of continually searching for the best methods, practices, and processes, and adopting or adapting the good features to become "the best of the best."

Blocking A situation that occurs because buffers have only limited capacity. If an output buffer is filled, processing at the upstream resource must halt because there is no place to store the already processed units. Similarly, if an input buffer gets filled, no more arrivals can be accommodated.

Bottleneck Slowest resource pool. See also *theoretical bottleneck*.

Buffer The part of the process that stores flow units that have finished with one activity but are waiting for the next activity to start.

Buffer capacity The maximum number of flow units that can wait in a buffer.

Bullwhip effect The phenomenon of upstream variability magnification that indicates a lack of synchronization among supply chain members.

Business process A network of activities performed by resources that transform inputs into outputs.

Business strategy The aspect of strategic planning that defines the scope of each division or business unit in terms of the attributes of the products that it will offer and the market segments that it will serve.

***c* chart** A chart that shows *control band* of acceptable variability in the number of defective flow units produced; also called *number of defects chart*.

Capacity utilization of a resource pool The degree to which resources are utilized by a process; the ratio of throughput and effective capacity of resource pool.

Capacity waste factor Fraction of theoretical capacity of a resource unit wasted.

Capital Fixed assets, such as land, building, facilities, equipment, machines, and information systems.

Cascading Representing a given process at several levels of detail simultaneously in a process flowchart.

Causal models Forecasting methods that assume data plus other factors influence demand.

Cause–effect diagram An illustration that shows a chain of cause–effect relationships that allows one to find the root causes of the observed variability. Also called *fishbone diagram* or *Ishikawa diagram.*

Cellular layout A layout of resources where all stations that perform successive operations on a product (or product family) are grouped together and organized according to the sequence of activities.

Changeover The cleaning, resetting, or retooling of equipment in order for it to process a different product. Also called *setup.*

Chase demand strategy A strategy to deal with demand fluctuations whereby a firm produces quantities to exactly match demand.

Check sheet A tally of the types and frequency of problems with a product or a service experienced by customers.

Coefficient of variation A measure of variability relative to its mean. It is obtained by computing the ratio of the standard deviation to the mean.

Collaborative Planning, Forecasting, and Replenishment (CPFR) An initiative in the consumer-goods industry designed to coordinate planning, forecasting, and replenishment across the supply chain.

Competitive product space A representation of the firm's product portfolio as measured along four dimensions or product attributes: product cost, response time, variety, and quality.

Continuous improvement Ongoing incremental improvement in process performance. See also *Kaizen*.

Continuous Replenishment Program (CRP) A partnership program under which the supplier automatically replenishes its customer inventories based on contractually agreed-on levels.

Continuous review, reorder point policy An order policy wherein a process manager, having initially ordered a fixed quantity, monitors inventory level continuously and then reorders once available inventory falls to a prespecified reorder point.

Control band A range within which any variation is to be interpreted as a normal, unavoidable aspect of any process.

Control chart A run chart of process performance with control limits overlaid to give it decision-making power.

Control limits The lower and upper range of the *control band*.

Corporate strategy The aspect of strategic planning that defines the businesses in which the corporation will participate and specifies how key corporate resources will be acquired and allocated to each business.

Critical activities Activities that lie on the critical path.

Critical path The longest path in the flowchart.

Cycle inventory The average inventory arising from a specific batch size.

Cycle service level The probability that there will be no stockout within a time interval. Also called *service level.*

Delayed differentiation The practice of reorganizing a process in order to delay the differentiation of a generic product to specific end-products closer to the time of sale. Also called *postponement.*

Demand management strategies Actions by a firm that attempt to influence demand pattern.

Division of labor The breakdown of labor into its components and the distribution of labor among people and machines to increase efficiency of production.

Economic order quantity The optimal order size that minimizes total fixed and variable costs.

Economies of scale A process exhibits economies of scale when the average unit cost of output decreases with volume.

Effective capacity of a process The effective capacity of the bottleneck.

Effective capacity of a resource unit The maximum flow rate of a resource unit if it were to be observed in isolation. It is equal to the inverse of the unit load.

Effective capacity of a resource pool Sum of the effective capacities of all the resource units in that pool.

Enterprise resource planning systems Information technology platform to gather and monitor information regarding materials, orders, schedules, finished goods inventory, receivables and other business processes across a firm.

Everyday low pricing (EDLP) The retail practice of charging constant, everyday low prices with no temporary discounts.

Everyday low purchase prices (EDLPP) The wholesale practice of charging constant, everyday low prices with no temporary discounts.

Feedback control The process of periodically monitoring the actual process performance, comparing it to planned levels of performance, investigating causes of the observed discrepancy between the two, and taking corrective actions to eliminate those causes.

Fill rate The fraction of total demand satisfied from inventory on hand.

Fishbone diagram See *cause–effect diagram*.

Fixed order cost The administrative cost of processing an order, transporting material, receiving the product(s), and inspecting the delivery regardless of order size.

Fixed setup cost The time and materials required to set up a process.

Flexible manufacturing system (FMS) A reprogrammable manufacturing system capable of producing a large variety of parts.

Flexible mass production A method of high-volume production that allows differences in products.

Flow rate The number of flow units that flow through a specific point in the process per unit of time.

Flow shop A type of process architecture that uses specialized resources to produce a low variety of products at high volumes.

Flow time The total time that a flow unit spends within process boundaries.

Flow-time efficiency The ratio between theoretical flow time and the average flow time that indicates the amount of waiting time associated with the process.

Flow unit The item being analyzed within a process view. Examples of flow units include an input unit, such as a customer order, or an output unit, such as a finished product. A flow unit can also be the financial value of the input or output.

Focused process A process in which products all fall within a small region of the competitive product space. This process supports a focused strategy.

Focused strategy A business process that is committed to a limited, congruent set of objectives in terms of demand and supply.

Forecasting The process of predicting the future.

Forward buying The taking advantage of price discounts to purchase for future needs.

Fraction defective chart See *p chart*.

Functional layout A type of process design that groups organizational resources by processing activity or "function" in "departments." Also called *process layout.*

Functional specialization A process and organizational structure where people are specialized by function, meaning each individual is dedicated to a specific task. (The other organizational structure is *product specialization.*)

Functional strategy The part of strategic planning that defines the purpose of marketing, operations, and finance—the three main functions of organizations.

Heijunka See *level production*.

Histogram A bar plot that displays the frequency distribution of an observed performance characteristic.

Ideal process A process that achieves synchronization at the lowest possible cost.

Incremental unit discount policy A quantity discount policy where a buyer receives discount only on units purchased above a certain threshold value; also see *all-unit quantity discount policy*.

Inflow rate The average rate of flow unit arrivals per unit of time.

Information technology Hardware and software used throughout businesses processes to support data gathering, planning, and operations.

In-process inventory An inventory classification; flow units that are being processed. See *work-in-process inventory* and *in-transit inventory*.

Inputs Any tangible or intangible items that flow into the process from the environment and are transformed; they include raw materials, component parts, energy, data, and customers in need of service.

Inputs inventory An inventory classification; flow units that are waiting to begin processing.

Instantaneous inventory accumulation rate The difference between instantaneous inflow rate and outflow rate, written $\Delta R(t)$. Also called *instantaneous inventory buildup rate.*

Instantaneous inventory buildup rate See *instantaneous inventory accumulation rate.*

Interarrival time The time between consecutive customer arrivals.

In-transit inventory A category of in-process inventory; flow units being transported. Also called *pipeline inventory.*

Inventory The total number of flow units present within process boundaries.

Inventory buildup diagram An illustration that depicts inventory fluctuation over time.

Inventory holding cost The financial cost of carrying inventory. Its two main components are *physical holding cost* and the *opportunity cost* of capital.

Inventory level The inventory on-hand.

Inventory position Sum of on-hand inventory and on-order inventory.

Inventory turns The ratio of throughput to average inventory. It is the reciprocal of average flow time. Also called *turnover ratio.*

Ishikawa diagram See *cause–effect diagram*.

Jidoka Intelligent automation whereby the ability to detect errors is automatically built into the machine.

Job shop A type of process architecture that uses flexible resources to produce low volumes of customized, high-variety products.

Just-in-time An action taken only when it becomes necessary to do so. In the context of manufacturing, it means production of only necessary flow units in necessary quantities at necessary times.

Kaizen Ongoing improvement of processes by continuously identifying and eliminating sources of waste in a process, such as inventory, waiting time, or defective parts.

Kanban The signaling device formalized by Toyota that allows the customer to inform the supplier of its need. It is a card attached to an output flow unit in the buffer between customer and supplier processes and lists the customer process, the supplier process, parts description, and production quantity.

Labor Human resource assets, such as engineers, operators, customer service representatives, and sales staff.

Lead time The time lag between the arrival of the replenishment and the time the order was placed.

Leadtime demand The total flow unit requirement during replenishment lead time.

Level production A production schedule where small quantities of different products are produced frequently to match with customer demand. Also called *heijunka.*

Level-production strategy The maintenance of a constant processing rate when demand fluctuates seasonally and thus the building of inventories in periods of low demand and the depleting of inventories when demand is high.

Little's law The law that describes the relationship among the flow time, inventory, and throughput. It states that average inventory equals average throughput times average flow time.

Load batching The phenomenon of a resource processing several flow units simultaneously; the number of units processed simultaneously is called the load batch.

Lot size The number of units processed consecutively after a setup. Also called *setup batch.*

Lower control limit (LCL) Lower range of the *control limits.*

Lower specification (LS) Lower range of acceptable performance.

Make-to-order Produce in response to customer orders.

Make-to-stock Produce in anticipation of customer orders.

Manufacturing The process of producing goods. Also called *product operations.*

Marginal analysis The process of comparing expected costs and benefits of purchasing each incremental unit.

Market-driven strategy One of two approaches to strategic fit wherein a firm starts with key competitive priorities and then develops processes to support them. (The other approach to strategic fit is *process-driven strategy.*)

Mass production The production of products in large (massive) quantities.

Material requirements planning (MRP) A planning tool in which the end-product demand forecasts are "exploded" backward to determine parts requirements at intermediate stations based on the product structure ("bill of materials"), processing lead times, and levels of inventories at those stations.

Multi-vari chart A plot of high, average, and low values of performance measurement sampled over time.

Net marginal benefit The difference between the unit price of the product and unit marginal cost of procurement.

Net marginal cost The difference between unit marginal cost of procurement and its salvage value.

Net present value A measure of expected aggregate monetary gain or loss that is computed by discounting all expected future cash inflows and outflows to their present value.

Network of activities and buffers Process activities linked so that the output of one becomes an input into another, often through an intermediate buffer.

Newsvendor problem A basic model of decision making under uncertainty whereby the decision maker balances the expected costs of ordering too much with the expected costs of ordering too little to determine the optimal order quantity.

Non-value-adding activities Activities that are required by a firm's process that do not directly increase the value of a flow unit.

Normal variability Statistically predictable variability. It includes both structural variability and *stochastic variability.*

Number of defectives chart See *c chart*.

On-order inventory The total inventory represented by all outstanding orders not yet delivered.

Operational effectiveness The measure of how well a firm manages its processes.

Operations Business processes that design, produce, and deliver goods and services.

Operations frontier The smallest curve that contains all industry positions in the competitive product space.

Operations strategy The aspect of strategic planning that configures and develops business processes that best enable a firm to produce and deliver the products specified by the business strategy.

Opportunity cost The forgone return on the funds invested in a given activity rather than in alternative projects.

Order upto level Target inventory level in a periodic review policy.

Outputs Any tangible or intangible items that flow from the process back into the environment. Examples include finished products, processed information, material, energy, cash, and satisfied customers.

Outputs inventory An inventory classification; processed flow units that have not yet exited process boundaries.

***p* chart** A chart shows *control band* of acceptable variability in the fraction of defective items produced; also called *fraction defective chart*.

Pareto chart A bar chart that plots frequencies of problem-type occurrence in decreasing order.

Periodic review policy An order policy wherein a process manager monitors inventory level periodically and reorders to bring the inventory position to a pre-determined order upto level; also called *periodic review order upto policy.*

Physical centralization The consolidation of all of a firm's stock into one location from which it services all customers.

Physical holding cost The out-of-pocket expense of storing inventory.

Pipeline inventory A category of in-process inventory; flow units being transported. Also called *in-transit inventory.*

Plan-Do-Check-Act (PDCA) cycle A tool to implement continuous improvement. It involves planning the process, operating it, inspecting its output, and adjusting it in light of the observation.

Plant Any singly owned, independently managed and operated facility, such as a manufacturing site, a service unit, or a storage warehouse.

Plant-within-a-plant (PWP) A plant in which the entire facility is divided into several "miniplants," each devoted to its own specific mission by performing a process that focuses strictly on that mission.

Poka yoke Mistake-proofing; design of a part, product, or a process that prevents its user from making a mistake.

Pooling capacity The sharing of available capacity among various sources of demand (or arrivals).

Pooling inventory The sharing of available inventory among various sources of demand.

Postponement See *delayed differentiation*.

Precedence relationships The sequential relationships that determine which activity must be finished before another can begin.

Process Any organization or *any part* of an organization that transforms inputs into outputs.

Process capability The ability of a process to meet customer specifications.

Process capacity The maximum sustainable flow rate of a process.

Process control The aspect of process management that is focused on continually ensuring that, in the short run, the actual process performance conforms to the planned performance.

Process cost The total cost incurred in producing and delivering outputs.

Process design The system of selecting the process architecture that best develops the competencies that will meet customer expectations.

Process efficiency Process performance measured in terms of total processing cost.

Process flexibility The ability of the process to produce and deliver desired product variety.

Process flowchart A graphical representation of a process that identifies the inputs, outputs, flow units, network of activities and buffers, resources allocated to activities, and information structure.

Process flow management A set of managerial policies that specifies how a process should be operated over time and which resources should be allocated over time to the activities.

Process flow measures Internal measures of process performance that managers can control. Together, these measures—flow time, flow rate, and inventory—capture the essence of process flow.

Process flow time See *flow time*.

Process layout See *functional layout*.

Process metrics Measurable dimensions along which the performance of the process will be tracked.

Process planning Identifying internal measures that track process competence and specifying the managerial policies that improve process competence along desired dimensions.

Process quality The ability of the process to produce and deliver quality products.

Process synchronization The ability of the process to meet customer demand in terms of their quantity, time, quality, and location requirements.

Process-driven strategy One of two approaches to strategic fit wherein a firm starts with a given set of process competencies and then identifies which market position is best supported by those processes. (The other approach to strategic fit is *market-driven strategy*.)

Processing network A system that consists of information and material flows of multiple products through a sequence of interconnected paths.

Processing rate The rate at which customers are processed by a server. Also called *service rate.*

Processing time See *activity time*.

Procurement batch A *batch* size arising in procurement.

Product attributes The properties of a product that customers consider important.

Product cost The total cost that a customer incurs in order to own and experience the product.

Product delivery response time The total time that a customer must wait for, before receiving a product for which he or she has expressed a need to the provider.

Product layout A type of process design in which the location of resources is dictated by the processing requirements of the product.

Product quality The degree of excellence of a product; how well a product performs.

Product specialization A process and organizational structure where people are specialized by product, meaning each individual is dedicated to a specific product line. The other organizational structure is called *functional* (or process) *specialization*.

Product value The maximum price a specific customer is willing to pay for a product.

Product variety The range of choices offered to the customer to meet his or her needs.

Production batch A *batch* size arising in production.

Productivity dilemma The choice between manufacturing goods with higher variety at a lower rate of productivity or manufacturing goods with lower variety at a higher rate of productivity.

Product–process matrix A tool used to match processes to products proposed by Hayes and Wheelwright (1979).

Promised duration The practice of promising a time frame within which the product will be delivered after an order has been placed.

Proportion abandoning The number of customers who enter the process but abandon it before being served.

Proportion blocked The average fraction of arrivals blocked from entering the process because the input buffer is full.

Pull A process where the signal to produce is triggered by the customer so that each station produces only on demand from its customer station.

Push A process where input availability, as opposed to customer need, triggers production.

Quality function deployment (QFD) A conceptual framework that can be used to translate customers' functional requirements of a product into concrete design specifications.

Quality of conformance How well the actual product conforms to the chosen design specifications.

Quality of design How well product specifications aim to meet customer requirements.

Quantity discount policy A pricing policy where prices depend on the quantity purchased; see *all-unit quantity discount policy* and *incremental unit discount policy*.

Queue length formula A formula for the average queue length as a function of the utilization, number of servers, and variability.

***R* chart** A chart that shows the *control band* of acceptable variability in sample ranges over time.

Rate of return The reward that an investor demands for accepting payment delayed by one period of time.

Reengineering Fundamental rethinking and radical redesign of business processes in order to achieve dramatic improvements in critical measures of performance, such as cost, quality, service, and speed.

Reorder point Inventory position at the time a reorder is placed in a continuous review policy.

Resource availability loss A category of factors that affect process capacity in which the resource itself is not available for processing.

Resource breakdown The unavailability of a resource for processing due to equipment malfunctioning.

Resource idleness A category of factors that affect process capacity in which the resource is available but is not processing units. Examples of factors include starvation and blocking.

Resource pool A collection of interchangeable resources that can perform an identical set of activities.

Resource pooling Making separate resource pools flexible to handle tasks performed by each other.

Resource unit Each unit in a resource pool.

Resources Tangible assets that help transform inputs to outputs in a process. They are usually divided into two categories: capital, which includes fixed assets such as land, building, facilities, equipment, machines, and information systems, and labor, which includes people such as engineers, operators, customer service representatives, and sales staff.

Return on total assets A common financial measure that shows how well a firm uses its assets to earn income for the stakeholders who are financing it.

Review period Length of time period a process manager chooses to review inventory position and take action.

Robust design The designing of a product in such a way that its actual performance will not suffer despite any variability in the production process or in the customer's operating environment.

Run chart A plot of some measure of process performance monitored over time.

Safety capacity The excess processing capacity available to handle customer inflows.

Safety inventory Inventory maintained to insulate the process from disruptions in supply or uncertainty in demand. Also called *safety stock.*

Safety stock See *safety inventory*.

Safety time The time margin that should be allowed over and above the expected time to deliver service in order to ensure that a firm will be able to meet the promised date.

Scatter plot A graph showing how a controllable process variable affects the resulting product characteristic.

Scheduled availability The amount of time that a resource unit is scheduled for operation.

Seasonal inventory Inventory that act as buffers to absorb seasonal fluctuations of supply and demand.

Service order discipline The sequence in which waiting customers are served.

Service operations Processes that deliver services.

Service rate See *processing rate*.

Setup See *changeover*.

Setup batch See *lot size*.

Single minute exchange of dies (SMED) A system by which the changeover times can be reduced to less than ten minutes.

Single-phase service process A service process in which each customer is processed by one server and all tasks performed by that server are combined into a single activity.

Six-sigma A process that produces only 3.4 defective units per million opportunities.

Slack time of an activity The extent to which an activity could be delayed without affecting process flow time.

Square root law The law states that the total safety inventory required to provide a specified level of service increases by the square root of the number of locations in which it is held.

Stability condition The requirement that the average inflow rate should be strictly less than the average processing rate to ensure a *stable process.* It is necessary to limit delays or queues.

Stable process A process in which, in the long run, the average inflow rate is the same as the average outflow rate.

Starvation The forced idleness of resources due to the unavailability of necessary inputs.

Statistical quality control A management approach that relies on sampling of flow units and statistical theory to ensure the quality of the process.

Stochastic variability The unpredictable or random variability experienced by service processes.

Strategic fit Having consistency between the competitive advantage that a firm seeks and the process architecture and managerial policies that it uses to achieve that advantage.

Subprocess Activities that are exploded into a set of subactivities that is then considered a process in its own right, with its own set of inputs, outputs, activities, and so forth.

Supply chain An entire network of interconnected facilities of diverse ownership with flows of information and materials between them. It can include raw materials suppliers,

finished-goods producers, wholesalers, distributors, and retailers.

Takt time The maximal time that each process resource can devote to a flow unit to keep up with the demand. It equals the available total processing time divided by total demand during that time. Considering only available operating time, takt time equals the reciprocal of throughput.

Theoretical bottleneck A resource pool with minimum theoretical capacity.

Theoretical capacity of a process The theoretical capacity of the bottleneck.

Theoretical capacity of a resource unit The maximum sustainable flow rate of the resource unit if all waste were eliminated (without idle periods, resource availability loss, and time lost to setups).

Theoretical capacity utilization of the resource pool The ratio of throughput and theoretical capacity. This is the maximal utilization one could achieve for a resource pool.

Theoretical flow time The minimum amount of time required for processing a typical flow unit without any waiting.

Theoretical inventory The minimum amount of inventory necessary to maintain a process throughput rate in equilibrium.

Throughput See *average flow rate*.

Throughput delay curve A graph displaying the average flow time of a process as a function of capacity utilization.

Throughput Improvement Mapping A process by which a process manager identifies the most likely source of additional throughput.

Time-series analyses Forecasting methods that rely solely on past data.

Total Quality Management (TQM) A management system that emphasizes holistic nature of quality starting with customer focus, building in quality, involving suppliers and employees, emphasizing prevention, early detection and correction, and continuous improvement over a long term.

Total unit load Total amount of time required by the resource unit to process one flow unit including allocated setup time based on the lot size of production.

Trade promotion A form of price discount wherein a discount is offered for only a short period of time.

Trade-off On the operations frontier, a decreasing of one aspect to increase another.

Transfer batch A *batch* size arising in transportation or movement.

Turnover ratio See *inventory turns*.

Type I error A situation when process performance falls outside the control band even with normal variability.

Type II error A situation when process performance measure falls within the control band, even though there is an assignable cause of abnormal variability.

Unit load of a resource unit Average amount of time required by the resource unit to process one flow unit, given the way the resource is utilized by the process.

Upper control limit (UCL) Upper range of the *control limits.*

Upper specification (US) Upper range of acceptable performance.

Value-adding activities Those activities that increase the economic value of a flow unit because the customer values them.

Value stream mapping A tool used to map the network of activities and buffers in a process identifying the activities that add value and those like waiting that are wasteful.

Vendor managed inventory (VMI) A partnership program under which the supplier decides the inventory levels to be maintained at its customer locations and arranges for replenishments to maintain these levels.

Virtual centralization A system in which inventory pooling in a network of locations is facilitated using information regarding availability of goods and subsequent transshipment of goods between locations to satisfy demand.

Waste The failure to match customer demand most economically by, for example, producing inefficiently, producing defective products, producing in quantities too large or too small, and delivering products too early or too late.

Work content of an activity The activity time multiplied by the average number of visits at that activity. It measures the total amount of time required to perform an activity during the transformation of a flow unit.

Work-in-process inventory A category of in-process inventory; flow units being processed in a manufacturing or service operation.

***X*-bar chart** A chart that shows the control band of acceptable variability in sample averages over time.

INDEX

Page references followed by *t* or *f* denote tables or figures.