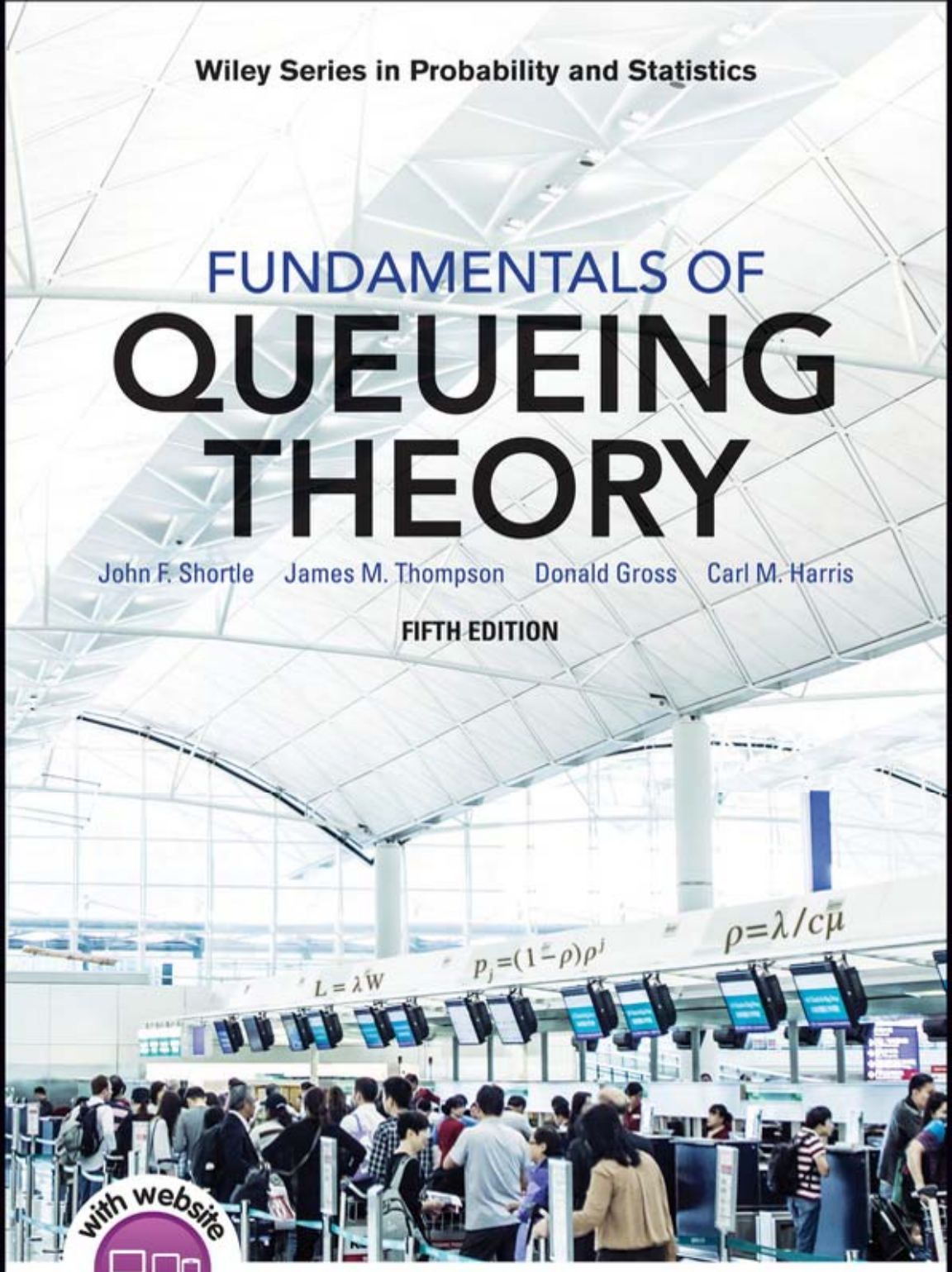
Wiley Series in Probability and Statistics

FUNDAMENTALS OF QUEUEING THEORY

John F. Shortle James M. Thompson Donald Gross Carl M. Harris

FIFTH EDITION



WILEY

FUNDAMENTALS OF QUEUEING THEORY

WILEY SERIES IN PROBABILITY AND STATISTICS
Established by *Walter A. Shewhart and Samuel S. Wilks*

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at
http://www.wiley.com/go/wsps

FUNDAMENTALS OF QUEUEING THEORY

FIFTH EDITION

JOHN F. SHORTLE
Professor of Systems Engineering & Operations Research
George Mason University

JAMES M. THOMPSON
Enterprise Architect
Freddie Mac

DONALD GROSS
Formerly of
George Mason University
Professor Emeritus
The George Washington University

CARL M. HARRIS
Late of
George Mason University

WILEY

This edition first published 2018

Edition History
John Wiley and Sons, Inc. (1e, 1974); John Wiley and Sons, Inc. (2e, 1985); Wiley-Interscience (3e, 1998); John Wiley and Sons, Inc. (4e, 2008)

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com. Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Shortle, John F., 1969- author. | Thompson, James M., 1954- author. | Gross, Donald, author. | Harris, Carl M., 1940-2000 author.

Title: Fundamentals of queueing theory / John F. Shortle, James M. Thompson, Donald Gross, Carl M. Harris.

Description: Fifth edition. | Hoboken, New Jersey : John Wiley & Sons, 2017. | Series: Wiley series in probability and statistics | Includes bibliographical references and index. |

Identifiers: LCCN 2017031755 (print) | LCCN 2017041116 (ebook) | ISBN 9781118943564 (pdf) | ISBN 9781118943533 (epub) | ISBN 9781118943526 (cloth) Subjects: LCSH: Queuing theory. Classification: LCC T57.9 (ebook) | LCC T57.9 .S54 2017 (print) | DDC 519.8/2--dc23 LC record available at https://lccn.loc.gov/2017031755

Cover image: ©RyanJLane/Gettyimages
Cover design by Wiley

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

The first edition of *Fundamentals of Queueing Theory*, written by Donald Gross and Carl Harris, was published in 1974. Since then, a new edition has appeared approximately once every ten years. In 2005, Donald Gross invited us (John Shortle and James Thompson) to help with a new edition, and we appreciate the opportunity to continue updating this excellent work. The changes in the fifth edition reflect the feedback from numerous students and colleagues since the fourth edition. Almost all of the material from the fourth edition has been kept, but with a fair amount of editing and reorganization. Several new sections have been added. We hope that the changes continue to bring improvements to the text.

One major change is that the first chapter from the fourth edition has been expanded and split into two chapters. The new Chapter 1 contains introductory material specific to queueing theory, while the new Chapter 2 contains general material on stochastic processes. In Chapter 1, a key addition is an expanded and more prominent section on Little's law. The treatment is more rigorous with multiple examples, a geometric proof, and extensions including the distributional form of Little's law and $H = \lambda G$. Chapter 1 also contains a new section on the psychology of waiting. In Chapter 2, the material on stochastic processes is rewritten and reorganized substantially from the fourth edition. The reorganization makes it more natural for someone who has covered the material elsewhere to skip the chapter. And for a reader who is

less familiar with the material, the chapter provides a concise treatment of essential results that are used throughout the text.

The chapter on advanced Markovian models (now Chapter 4) has been edited substantially and contains a new section on fairness in queueing as well as a discussion of processor sharing. The chapter on bounds and approximations (now Chapter 8) includes a new section on fluid queues. Many new examples and problems have been added throughout the text (over 20 new examples and over 60 new problems). Finally, the QtsPlus software has been been updated to run on the latest versions of Excel for both PCs and Macs. The user interface has also been improved significantly.

For errata, updates, and other information about the text and associated QtsPlus software, see the text website:

<http://mason.gmu.edu/~jshortle/fqt5th.html>.

John F. Shortle
James M. Thompson

Fairfax, Virginia
October 2017

ACKNOWLEDGMENTS

We are grateful for the opportunity participate in the writing of the fourth and fifth editions and acknowledge the enormous amount of work carried out by the original authors, Donald Gross and Carl Harris, in writing the first three editions. We humbly acknowledge that we stand on the shoulders of giants and hope that the changes made in the recent edition continue to improve the quality of the textbook.

We are grateful for the assistance given to us by many professional colleagues and students whose numerous comments and suggestions have been so helpful in improving this text. With heartfelt thanks, we extend special appreciation to our families for their unlimited and continuing encouragement and to all the people at John Wiley & Sons who have been wonderfully supportive. John also appreciates the support of the Volgenau School of Engineering and the Department of Systems Engineering and Operations Research at George Mason University.

J. F. S.
J. M. T.

ABOUT THE COMPANION WEBSITE

This book is accompanied by a companion website:

www.wiley.com/go/shortle/queueingtheory5e

The Student's website includes:

- A partial Solutions Manual

The Instructor's website (password protected with ProfVal Validation) includes:

- A complete Solutions Manual
  - To gain access to the site, instructors should follow instructions from the above link.

CHAPTER 1

INTRODUCTION

All of us have experienced the annoyance of having to wait in line. Unfortunately, this phenomenon continues to be common in congested, urbanized, "high-tech" communities. We wait in line in our cars in traffic jams or at toll booths; we wait on hold for an operator to pick up our telephone calls; we wait in line at supermarkets to check out; we wait in line at fast-food restaurants; and we wait in line at stores and post offices. We, as customers, do not generally like these waits, and the managers of the establishments at which we wait also do not like us to wait, since it may cost them business. Why then is there waiting?

The answer is simple: There is more demand for service than there is facility for service available. Why is this so? There may be many reasons; for example, there may be a shortage of available servers, it may be infeasible economically for a business to provide the level of service necessary to prevent waiting, or there may be a space limit to the amount of service that can be provided. Generally these limitations can be removed with the expenditure of capital, and to know how much service should then be made available, one would need to know answers to such questions as "How long must a customer wait?" and "How many people will form in the line?" Queueing theory attempts to answer these questions through detailed mathematical analysis.

Fundamentals of Queueing Theory, Fifth Edition.
By J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris

The earliest problems studied in queueing theory were those of telephone traffic congestion. The pioneer investigator was the Danish mathematician A. K. Erlang, who, in 1909, published "The Theory of Probabilities and Telephone Conversations." In later works he observed that a telephone system was generally characterized by either (1) Poisson input, exponential holding (service) times, and multiple channels (servers), or (2) Poisson input, constant holding times, and a single channel. Work on the application of the theory to telephony continued after Erlang. In 1927, E. C. Molina published his paper "Application of the Theory of Probability to Telephone Trunking Problems," which was followed one year later by Thornton Fry's book *Probability and Its Engineering Uses*, which expanded much of Erlang's earlier work. In the early 1930s, Felix Pollaczek did some further pioneering work on Poisson input, arbitrary output, and single- and multiple-channel problems. Additional work was done at that time in Russia by Kolmogorov and Khintchine, in France by Crommelin, and in Sweden by Palm. The work in queueing theory picked up momentum rather slowly in its early days, but accelerated in the 1950s, and there has been a great deal of work in the area since then.

There are many valuable applications of queueing theory including traffic flow (vehicles, aircraft, people, communications), scheduling (patients in hospitals, jobs on machines, programs on a computer), and facility design (banks, post offices, amusement parks, fast-food restaurants). Most real problems do not correspond exactly to a mathematical model, and increasing attention is being paid to complex computational analysis, approximate solutions, simulation, and sensitivity analyses.

1.1 Measures of System Performance

Figure 1.1 shows a typical queueing system: Customers arrive, wait for service, receive service, and then leave the system. Some customers may leave without receiving service, perhaps because they grow tired of waiting in line or perhaps because there is no room to enter the service facility in the first place.

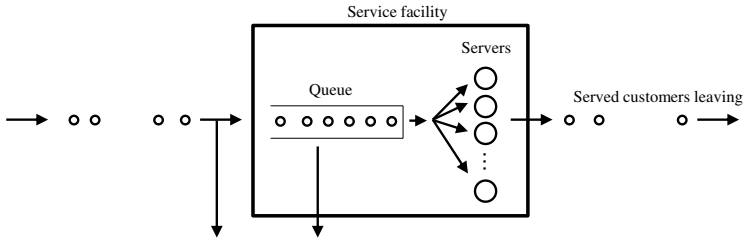


Figure 1.1 A typical queueing system.

Note that the term "customer" is often used throughout this text in a general sense and does not necessarily imply a human customer. For example, a customer could

be a ball bearing waiting to be polished, an airplane waiting in line to take off, or a computer program waiting to be run.

What might one like to know about the effectiveness of a queueing system? Generally there are three types of system responses of interest: (1) Some measure of the *waiting time* that a typical customer might endure, (2) some measure of the *number of customers* that may accumulate in the queue or system, and (3) a measure of the *idle time* of the servers. Since most queueing systems have stochastic elements, these measures are often random variables, so their probability distributions – or at least their expected values – are sought.

Regarding waiting times, there are two types – the time a customer spends in the queue and the total time a customer spends in the system (queue plus service). Depending on the system being studied, one may be of more interest than the other. For example, if we are studying an amusement park, it is the time waiting in the queue that makes the customer unhappy. But if we are dealing with machines that require repair, then it is the total down time (queue wait plus repair time) that we wish to keep as small as possible. Throughout this book, the average waiting time of a typical customer in queue is denoted as W_q and the average waiting time in the system is denoted as W.

Correspondingly, there are two customer accumulation measures – the number of customers in the queue and the total number of customers in the system. The former is of interest if we desire to determine a design for waiting space (e.g., the number of seats to have for customers waiting in a hair-styling salon), while the latter may be of interest for knowing how many machines may be unavailable for use. The average number of customers in the queue is denoted as L_q and the average number of customers in the system is denoted as L. Finally, idle-service measures can include the percentage of time any particular server may be idle or the time the entire system is devoid of customers.

The task of the queueing analyst is generally one of two things – to determine some measures of effectiveness for a given process or to design an "optimal" system according to some criterion. To do the former, one must determine waiting delays and queue lengths from the given properties of the input stream and the service procedures. For the latter, the analyst might want to balance customer-waiting time against the idle time of servers according to some cost structure. If the costs of waiting and idle service can be obtained directly, they can be used to determine the optimum number of servers. To design the waiting facility, it is necessary to have information regarding the possible size of the queue . There may also be a space cost that should be considered along with customer-waiting and idle-server costs to obtain the optimal system design. In any case, the analyst can first try to solve this problem by analytical means; if these fail, he or she may use simulation. Ultimately, the issue generally comes down to a trade-off between better customer service and the expense of providing more service capability, that is, determining the increase in investment of service for a corresponding decrease in customer delay.

1.2 Characteristics of Queueing Systems

A quantitative evaluation of a queueing system requires a mathematical characterization of the underlying processes. In many cases, six basic characteristics provide an adequate description of the system:

1. Arrival pattern of customers
2. Service pattern of servers
3. Number of servers and service channels
4. System capacity
5. Queue discipline
6. Number of service stages

The standard notation for characterizing a queueing system based on the first five characteristics will be described shortly (Section 1.2.7).

1.2.1 Arrival Pattern of Customers

In usual queueing situations, the process of arrivals is stochastic, and it is thus necessary to know the probability distribution describing the times between successive customer arrivals (interarrival times). A common arrival process is the *Poisson process*, which will be described in Section 2.2. It is also necessary to know whether customers can arrive simultaneously (batch or bulk arrivals), and if so, the probability distribution describing the size of the batch.

Another factor is the manner in which the pattern changes with time. An arrival pattern that does not change with time (i.e., the probability distribution describing the input process is time-independent) is called a *stationary* arrival pattern. One that is not time-independent is called *nonstationary*. An example of a system with a nonstationary arrival pattern might be a restaurant where more customers tend to arrive during the lunch hour than during other times of the day. Many of the models in this text assume a stationary arrival process.

It is also necessary to know the reaction of a customer upon arrival to the system. A customer may decide to wait no matter how long the queue becomes, or, if the queue is too long, the customer may decide not to enter the system. If a customer decides not to enter the queue upon arrival, the customer is said to have *balked*. A customer may enter the queue, but after a time lose patience and decide to leave. In this case, the customer is said to have *reneged*. In the event that there are two or more parallel waiting lines, customers may switch from one to another, that is, *jockey* for position. These three situations are all examples of queues with *impatient customers*.

1.2.2 Service Patterns

Much of the previous discussion concerning the arrival pattern is appropriate in discussing service. Most important, since service times are typically stochastic, a probability distribution is needed to describe the sequence of customer service times. Service may also be single or batch. One generally thinks of one customer being served at a time by a given server, but there are many situations where customers may be served simultaneously by the same server, such as a computer with parallel processing, sightseers on a guided tour, or people boarding a train. The service process may also depend on the number of customers waiting for service. A server may work faster if the queue is building up or, on the contrary, may get flustered and become less efficient. The situation in which service depends on the number of customers waiting is referred to as *state-dependent* service. Service, like arrivals, can be stationary or nonstationary with respect to time. For example, learning may take place, so that service becomes more efficient as experience is gained. The dependence on time is not to be confused with dependence on state. The former depends on how long the system has been in operation (regardless of the state of the system), while the latter depends on the number of customers in the system (regardless of how long the system has been in operation). Of course, a queueing system can be both nonstationary and state-dependent.

1.2.3 Number of Servers

The number of servers is an important characteristic of a queueing system and represents a fundamental trade-off – adding servers incurs extra cost to the business, but can substantially reduce delays for customers. Thus, the choice of the number of servers is often a critical decision. Section 3.4 describes a rule of thumb for the trade-off between the number of servers and the customer delays.

Another decision is the configuration of the lines. For a multiserver system, there are several possible configurations. Figure 1.2 illustrates two main cases. In the first case, the servers are fed by a single queue. An example might be a baggage-check counter for an airline. Another example might be a hair-styling salon with many chairs, assuming no customer is waiting for any particular stylist. In the second case, each server is fed by its own queue. A grocery store might be an example of this case. Hybrid situations can also occur. For example, a passport line at an airport might initially start as a long single line and then later split into short separate lines for each agent. As we explain later, it is generally preferable for a multiserver queueing system to be fed by a single line. Thus, when specifying the number of parallel servers, we typically assume that the servers are fed by a single line. Also, it is generally assumed that the servers operate independently of each other.

1.2.4 Queue Discipline

Queue discipline refers to the manner in which customers are selected for service when a queue has formed. A common discipline in everyday life is first come,

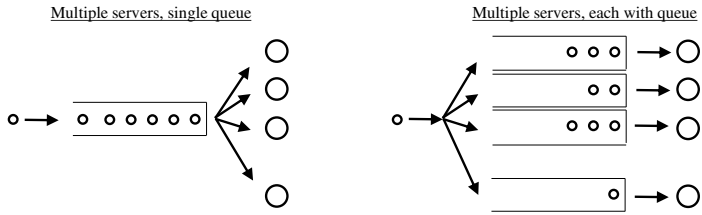


Figure 1.2 Multiserver queueing systems.

first served (FCFS). However, there are many other disciplines. Some other queue disciplines are: Last come, first served (LCFS), which is applicable to many inventory systems, as it is easier to reach the nearest items which are the last in; random selection for service (RSS) in which customers are selected randomly from the queue independent of their arrival times; processor sharing (PS) in which the server processes all customers (or jobs) simultaneously but works at a slower rate on each job based on the number in the system (this is common in computer systems); polling, in which a single server serves multiple queues by taking customers from the first queue, then customers from the second, and so forth in a cycle (a traffic light is a kind of polling system); and a variety of priority schemes where some customers receive preference in terms of being selected for service.

Priority schemes are treated in more detail in Section 4.4. In these disciplines, customers with higher priorities are selected for service ahead of those with lower priorities. There are two general situations in priority disciplines, *preemptive* and *nonpreemptive*. In the nonpreemptive case, the highest priority customer goes to the head of the queue but cannot get into service until the customer presently in service is completed, even if this customer has a lower priority. In the preemptive case, a higher priority customer is allowed to enter service immediately upon arrival even if a customer with lower priority is already in service. Service for the lower priority customer is interrupted, to be resumed again after the higher priority customer is served. There are two variations of the preemptive case: the preempted customer's service can either continue from the point of preemption or start anew.

1.2.5 System Capacity

In some systems, there is a physical limitation to the amount of space for customers to wait, so that when the line reaches a certain length, no further customers are allowed to enter until space becomes available. These are referred to as finite queueing situations; that is, there is a finite limit to the maximum system size. A queue with limited waiting room can be viewed as one where a customer is forced to balk if it arrives when the queue size is at its limit.

1.2.6 Stages of Service

A queueing system could have only a single stage of service, or it could have several stages. An example of a multistage queueing system is a physical examination procedure where each patient must proceed through several stages, comprising medical history; ear, nose, and throat examination; blood tests; electrocardiogram; eye examination; and so on. Multistage queueing processes are treated in Section 5.1, as a special case of more general queueing networks. In some multistage queueing processes, recycling or feedback may occur (Figure 1.3). Recycling is common in manufacturing processes, where quality control inspections are performed after certain stages, and parts that do not meet quality standards are sent back for reprocessing. Similarly, a telecommunications network may process messages through a randomly selected sequence of nodes, with the possibility that some messages will require rerouting through the same stage.

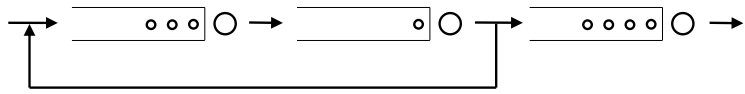


Figure 1.3 Multistage queueing system with feedback.

1.2.7 Notation

As shorthand for describing queueing processes, a notation has evolved, due for the most part to Kendall (1953), which is now rather standard throughout the queueing literature. A queueing process is described by a series of symbols and slashes $A/B/X/Y/Z$, where A denotes the interarrival-time distribution, B denotes the service-time distribution, X denotes the number of parallel servers, Y denotes the system capacity, and Z denotes the queue discipline. Table 1.1 presents some standard symbols for these characteristics (see also Appendix A for a dictionary of symbols and abbreviations used throughout the text).

For example, $M/D/2/\infty$/FCFS indicates a queueing system with exponential interarrival times, deterministic service times, two parallel servers, infinite system capacity (i.e., no restriction on the maximum number allowed in the system), and first-come, first-served queue discipline. In many situations only the first three symbols are used. Typical practice is to omit the service capacity if no restriction is imposed ($Y = \infty$) and to omit the queue discipline if it is first come, first served ($Z =$ FCFS). Thus $M/D/2$ would be the same as $M/D/2/\infty$/FCFS.

The symbols in Table 1.1 are, for the most part, self-explanatory; however, a few require further comment. First, it may appear strange that the symbol M is used for the exponential distribution. One might expect the use of the symbol E. However, this would be too easily confused with E_k, which is used for the Erlang distribution. Rather, M is used, standing for the Markovian or memoryless property of the exponential (described in Section 2.1). Second, the symbol G represents a general probability distribution. No assumption is made as to the precise form of

Table 1.1 Queueing notation $A/B/X/Y/Z$

Characteristic	Symbol	Explanation
Interarrival-time distribution (A)	M	Exponential
	D	Deterministic
	E_k	Erlang type k ($k = 1, 2, \ldots$)
Service-time distribution (B)	H_k	Mixture of k exponentials
	PH	Phase type
	G	General
Parallel servers (X)	$1, 2, \ldots, \infty$	
System capacity (Y)	$1, 2, \ldots, \infty$	
Queue discipline (Z)	FCFS	First come, first served
	LCFS	Last come, first served
	RSS	Random selection for service
	PR	Priority
	GD	General discipline

the distribution. Results in these cases are applicable to any probability distribution. Finally, the table is not complete. For example, there is no indication of a symbol to represent bulk arrivals or series queues. In many cases, the notation for a particular model is brought up when the model is introduced in the text. In some cases, there are models for which no symbolism has either been developed or accepted as standard, and this is generally true for models less frequently analyzed in the literature.

1.2.8 Model Selection

The six characteristics discussed in this section are sufficient to completely describe many queueing systems of interest. However, since a wide variety of queueing systems can be encountered in practice, it is critical to understand the system under study in order to select the model that best describes the real situation. A great deal of thought is often required in this *model selection procedure*, and knowledge of the six basic characteristics is essential in this task.

For example, consider the case of a supermarket. Suppose there are c checkout counters. If customers choose a checkout counter on a purely random basis (without regard to the queue length in front of each counter) and never switch lines (no jockeying), then we have c independent single-server models. If, instead, there is a single waiting line for all the counters, we have a c-server model with a single queue. Neither, of course, is generally the case in most supermarkets. What usually happens is that queues form in front of each counter, but new customers enter the queue that is the shortest (or has shopping carts that are lightly loaded). Also, there is a great deal of jockeying between lines. Now the question becomes which choice of models is

more appropriate. With jockeying, the c-server model with a single queue would be more appropriate. This is because a waiting customer always moves to a server that becomes idle. Thus, no server is idle while there are customers waiting for service. This behavior holds for the c-server queue but not for c independent single-server queues. As jockeying is rather easy to accomplish in supermarkets, the c-server model with one queue may be more appropriate and realistic than c independent single-server models, which one might have been tempted to choose initially prior to giving much thought to the process.

1.3 The Experience of Waiting

This textbook deals primarily with *quantitative* measures of waiting, such as W, W_q, L, and L_q. In this section, we give a brief interlude to mention some *qualitative* aspects of waiting. While a manager can improve quantitative measures of waiting by hiring more servers, the *experience* of waiting can also be improved in a number of other ways. This section summarizes several principles, proposed by Maister (1984), related to the experience or psychology of waiting. The reader can likely relate to many of these principles, recalling personal experiences when a given wait was more aggravating than it needed to be. See Maister (1984) for further discussion.

1. Unoccupied time feels longer than occupied time. If a customer can be kept busy while waiting, the delay does not feel as long. For example, a restaurant may hand out menus to waiting customers or may invite them to the bar. Moving the line in stages can also occupy time. For example, a sandwich shop may have multiple stages in line: Customers place their order with one server, choose sandwich toppings with another server, and finally pay with a third server. The gradual progress occupies time and reduces perceived wait.

2. Pre-process wait feels longer than in-process wait. Pre-process wait occurs before service starts, while in-process wait occurs after service starts. For example, when sitting down at a restaurant, if the server comes by and takes an initial drink order or says "I'll be with you in a moment," there is a perception that service has been initiated. The initial contact is important, and the wait prior to this contact may be perceived as longer.

3. Anxiety makes waiting seem longer. Anxiety can arise for a number of reasons. Am I in the wrong line? Will I be able to make my flight? Will I be able to board the next shuttle or will it be too crowded? Should I move to the other line that is moving faster? In some situations, anxiety can be reduced by having someone walk the lines explaining which line is which, assuring people that they will make their flight, and so forth.

4. Uncertain waits are longer than known, finite waits. A customer can often estimate the waiting time with a quick scan of the line length. However, when the line is very long or moving very slowly, it may be difficult to judge. Also, when the queue is virtual (e.g., a call center), there is no way to "see" the line. Providing an estimate of waiting time can reduce uncertainty for the customer. However, this also raises expectations. If the delay turns out to be longer than the estimate, this

may be more aggravating for the customer than providing no estimate. Conversely, overestimating the delay may unnecessarily turn customers away.

5. Unexplained waits are longer than explained waits. Customers are more patient if they know why a delay is occurring, particularly if the cause is viewed as justifiable (e.g., a thunderstorm that reduces airport capacity). In off-nominal situations, it can be helpful to make an announcement explaining the situation. However, a generic explanation ("We are currently experiencing a high volume of calls") may not be viewed as justifiable (Isn't there always a high volume of calls?).

6. Unfair waits are longer than equitable waits. One principle of fairness is that an earlier arriving customer should begin service before a later arriving customer (first come, first served, or FCFS). Situations that do not follow FCFS may be deemed unfair. For example, a grocery store may have separate lines for each server. While each line operates *individually* on a FCFS basis, the system as a whole may not. If the other line is moving faster, it becomes frustrating to see people who arrive after you begin service before you. Systems with no well-formed line can also be unfair. An example might be a shuttle stop where people gather as a nebulous group and board in somewhat random order. If the shuttle has limited space, the ones who are left to wait for the next shuttle are not necessarily the last to arrive. Priority-based systems (Section 4.4) violate FCFS and may or may not be viewed as fair. In an emergency room, it is accepted that medical emergencies receive service ahead of people with non-urgent needs. In other systems, priority service may be given to customers who pay a premium (fast pass lines at amusement parks), which may or may not be viewed as fair.

7. Longer waits are tolerable for more valuable service. Customers who receive longer service (which may correlate with the "value" of the service) may tolerate longer waits. For example, when purchasing a full cart of items at a grocery store, a longer wait may be more tolerable than when purchasing a single item. This raises a second principle of fairness – a customer with a shorter service time should wait less than a customer with a longer service time, all else being equal. This principle can be in tension with FCFS. What happens when a customer with a single item arrives behind a customer with a full cart of groceries? Should that customer be allowed to jump ahead? At a restaurant, is it acceptable to allow smaller groups to be seated ahead of larger ones? This tension and the issue of fairness will be discussed in more detail in Section 4.4.4.

8. Solo waits feel longer than group waits.

1.4 Little's Law

A fundamental relationship that is used extensively in queueing theory and throughout this text is *Little's law*. Little's law provides a relationship between three fundamental quantities: The average rate λ that customers arrive to a system, the average time W that a customer spends in the system, and the average number L of customers in the system. This relationship is given by $L = \lambda W$. Given two of the three quantities, one can infer the third. For example, if one is able to observe customers leaving

a store (yielding an estimate for λ) and one can ask each customer how long he or she was in the store (estimating W), then one can estimate L the average number of customers in the store.

Little's law is a very general result and can be applied to a wide variety of systems, even systems that might not be considered queues. Before stating the result formally, we give an example to illustrate the principle.

■ **EXAMPLE 1.1**

An elementary school has 6 grades (1st grade through 6th grade). Every year, 30 new students enroll in first grade. The students progress through the successive grades and leave upon completing 6th grade. What is the total number of students enrolled at the school?

The answer is straight-forward: The arrival rate to the system is $\lambda = 30$ new students per year. Each student remains in the school for 6 years, so $W = 6$. By Little's law, the total average enrollment in the school is $L = \lambda W = 180$.

This example illustrates that Little's law might be considered an "obvious" relationship. Each grade has 30 students. There are 6 grades. So the total number of students is 180. Yet this argument implicitly makes a number of assumptions. For example, the argument assumes that the students proceed in a deterministic manner through each grade. What if some students enter and/or leave at intermediate grades? What if some students skip or repeat grades? What if the enrollment numbers vary from year to year in a stochastic manner? What if the enrollment numbers slowly increase over time?

To address these questions more carefully, we now give a mathematically precise statement of Little's law. Consider a system with arriving and departing customers (Figure 1.4). Let $A^{(k)}$ be the time that customer k enters the system, where $A^{(k)}$ is ordered so that $A^{(k+1)} \geq A^{(k)}$. Let $A(t)$ denote the cumulative number of arrivals to the system by time t. Let $W^{(k)}$ be the time that customer k spends in the system. A customer cannot depart before arriving, so $W^{(k)} \geq 0$. Let $N(t)$ be the number of customers in the system at time t. That is, $N(t)$ is the number of indexes k such that $A^{(k)} \leq t$ and $A^{(k)} + W^{(k)} \geq t$. Define the following limits, when they exist:

$$\lambda \equiv \lim_{t\to\infty} \frac{A(t)}{t}, \quad W \equiv \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} W^{(k)}, \quad L \equiv \lim_{T\to\infty} \frac{1}{T}\int_0^T N(t)\,dt. \quad (1.1)$$

The first limit λ is the long-run average rate of arrivals. The second limit W is the long-run average time spent in the system per customer. The third limit L is the long-run average number of customers in the system.

Theorem 1.1 [Little's law] *If the limits λ and W in (1.1) exist and are finite, then the limit L exists and*

$$\boxed{L = \lambda W.}$$

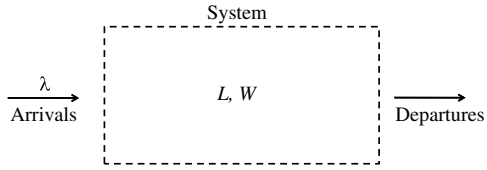


Figure 1.4 Generic setting for Little's law.

Proofs can be found, for example, in Stidham (1974) and Wolff (2011); a minor variant is proved in Whitt (1991). The relationship can also be proved with slightly different assumptions on the underlying stochastic processes. The original proof by Little in 1961 requires the underlying processes to be strictly stationary, as does the theorem in Brumelle (1971a). Some other versions require the existence of regeneration points when the system empties out and "starts over" (e.g., Jewell, 1967). Some variants of the theorem in finite time are given by Little (2011) in a retrospective article.

Before giving examples, we make some general remarks about Little's law. First, Theorem 1.1 is a statement about *long-run averages*. That is, the quantities L, λ, and W in (1.1) are all defined as *infinite limits*. Many of the results in this book are stated using infinite long-run averages, so Little's law provides necessary relationships in the derivation of this theory.

Second, Theorem 1.1 requires that the limits for λ and W exist. This precludes scenarios in which the time in system is growing without bound. This occurs in an unstable queue where the arrival rate exceeds the maximum service rate, so the queue size (and hence the time in the system) grows without bound over time.

Third, the theorem does not technically require the existence of a "queue." Rather, it requires the existence of a "system" to which entities arrive and from which they depart. The system can be regarded as a black box, and there are no specific requirements about what happens inside the black box, aside from the existence of appropriate limits as stated previously. For example, there is no requirement that entities depart in the order they arrive. There is no requirement of Poisson arrivals, exponential service, or FCFS service discipline (common assumptions throughout the text). The main requirement is that entities depart after they arrive (i.e., $W^{(k)} \geq 0$).

Depending on how the "system" is defined, different relationships can be derived from Little's law, as the following examples illustrate. In this sense, Little's law can be thought of as a principle, rather than a fixed equation. In particular, for a given queueing system, the quantities L, λ, and W can take on different meanings depending on how the system is defined with respect to the queue.

■ EXAMPLE 1.2

Figure 1.5 shows a common representation of Little's law. The system includes both the queue and the server. This is the typical meaning of "system" in this book. With this definition, L refers to the average total number of customers in the system, including customers in the queue and customers in service. W

refers to the total average time in the system, from the initial arrival time to the final departure time (time in queue plus time in service). Little's law then implies that $L = \lambda W$.

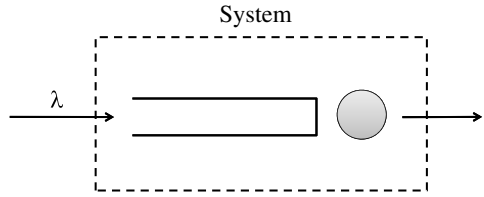


Figure 1.5 Little's law.

While Figure 1.5 shows a single queue and a single server, the same relationship holds if the system contains multiple servers and/or multiple queues.

EXAMPLE 1.3

Figure 1.6 considers the "system" as the queue. Little's law implies that

$$L_q = \lambda W_q,$$

where L_q is the average number of customers in the queue and W_q is the average time a customer spends in the queue. The arrival rate to the queue is the same as the arrival rate to the whole system (i.e., λ).

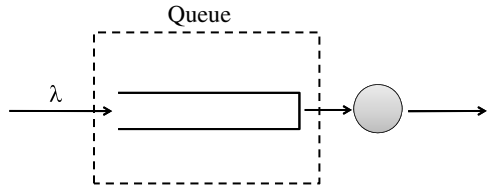


Figure 1.6 Little's law applied to the queue.

EXAMPLE 1.4

This example considers the "system" as the single server (Figure 1.7). In this case, L represents the average number of customers in service. Since there is only one server, the average number in service is $0 \cdot p_0 + 1 \cdot (1 - p_0) = 1 - p_0$, where p_0 is the fraction of time the system is empty. W represents the average time a customer spends in service, or $\mathrm{E}[S]$ where S is a random service time. Assuming a stable queue (i.e., where the long-run rate that customers leave the queue is the same as the long-run rate they enter the queue), the arrival rate to the server is λ. Thus, "$L = \lambda W$" becomes

$$1 - p_0 = \lambda \cdot \mathrm{E}[S]. \tag{1.2}$$

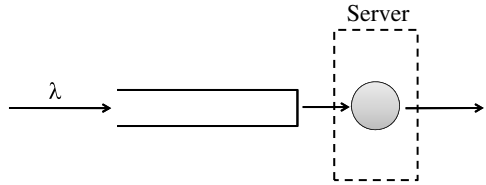


Figure 1.7 Little's law applied to the server.

This relationship has been derived under very general conditions. In particular, the equation does *not* require many of the common assumptions used elsewhere in this book, such as Poisson arrivals, exponential service, or a first-come, first-served service discipline. The equation does, however, require a *single* server. (For more than one server, the average number in service L is no longer $1 - p_0$, as it is for a single server.)

EXAMPLE 1.5

This example considers a queue with *blocking* (Figure 1.8). Blocking occurs in systems with finite capacity. An arriving customer who finds the system full is assumed to depart without entering the system. These models are common in telecommunications where the service provider has a finite capacity to handle incoming calls (e.g., see Sections 3.5 and 3.6). Suppose that a certain fraction p_b of arrivals is blocked and does not enter the system. Thus, the rate that customers enter the system is $(1 - p_b)\lambda$. Little's law yields

$$L = (1 - p_b)\lambda W.$$

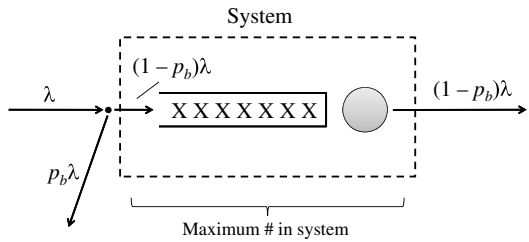


Figure 1.8 Little's law applied to a queue with blocking.

In this example, care must be taken in the interpretation of W. Since the blocked customers do not enter the system, these customers are not counted in the average for W. That is, W represents the average time spent in the system *among those customers who actually enter the system.*