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.

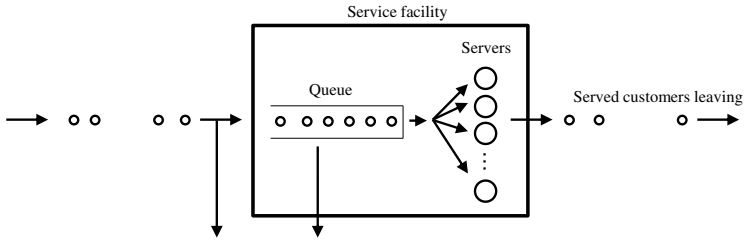


Figure 1.1 A typical queueing system.

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,

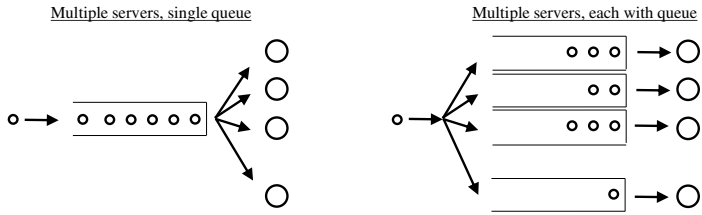


Figure 1.2 Multiserver queueing systems.

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.

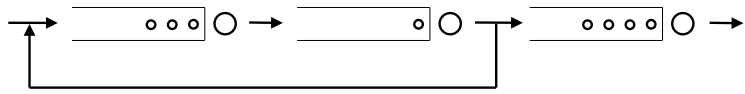


Figure 1.3 Multistage queueing system with feedback.

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*

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

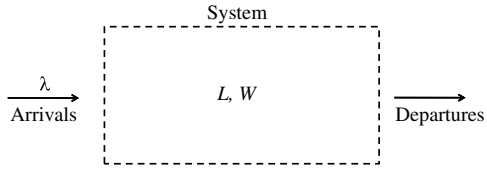


Figure 1.4 Generic setting for Little's law.

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 $L = \lambda W$.

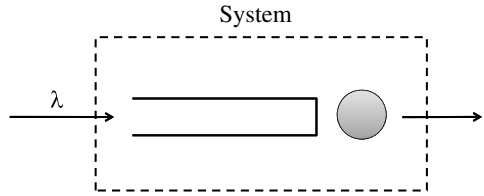


Figure 1.5 Little's law.

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).

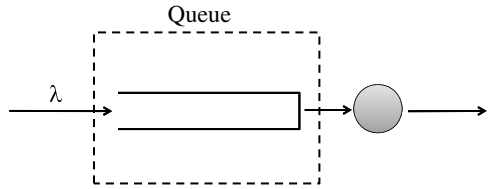


Figure 1.6 Little's law applied to the queue.

■ **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

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

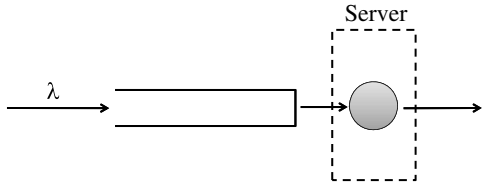


Figure 1.7 Little's law applied to the server.

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

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

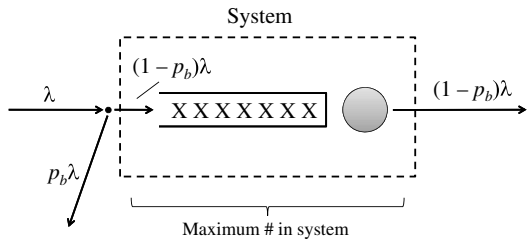


Figure 1.8 Little's law applied to a queue with blocking.

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.*

1.4.1 Geometric Illustration of Little's Law

We now give a geometric "proof" of Little's law. This is not a rigorous proof, but rather a rough argument showing the main ideas behind Little's law. Full technical proofs can be found in the references cited earlier. In the geometric argument, we consider a system that *starts and ends in an empty state*. We also assume that *customers depart in the order that they arrive*, though this assumption will be relaxed later.

Let $A(t)$ and $D(t)$ denote the cumulative number of arrivals and departures by time t. Figure 1.9 shows sample paths for $A(t)$ (solid line) and $D(t)$ (dashed line). The number of customers in the system at time t is $A(t)-D(t)$, so the system is empty whenever $A(t) = D(t)$. In this example, the system starts and ends in an empty state. There is also an intermediate point where the system empties temporarily. Let N denote the number of arrivals on the time horizon $[0, T]$; here, $N = 6$.

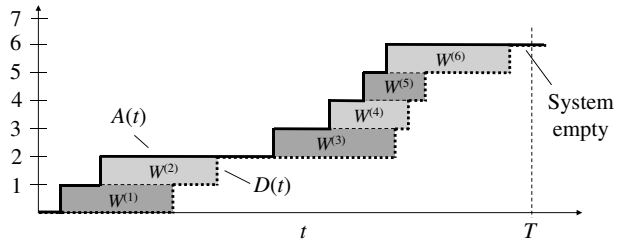


Figure 1.9 Geometric representation of Little's law (customers depart in order).

Because customers depart in the order of arrival, each horizontal rectangle represents the time a particular customer k spends in the system, namely $W^{(k)}$. The total time spent in the system among all customers, $W^{(1)} + W^{(2)} + \ldots + W^{(N)}$, is the total area of the rectangles.

Now, the shaded area can also be measured by integrating $A(t) - D(t)$ over $[0, T]$. So

$$\int_0^T A(t) - D(t)\, dt = \text{area of rectangles} = \sum_{k=1}^{N} W^{(k)}. \tag{1.3}$$

Dividing both sides by T gives

$$\frac{1}{T}\int_0^T A(t) - D(t)\, dt = \frac{N}{T} \cdot \left(\frac{1}{N} \sum_{k=1}^{N} W^{(k)} \right),$$

where we have also multiplied and divided by N on the right-hand side. The left-hand side is the time average of $A(t) - D(t)$, which is the average number of customers in the system, or L. On the right-hand side, the first term N/T represents the number of arrivals per time, or λ. The second term is the average time spent in the system per customer, or W. Thus, we have

$$L = \lambda W.$$

Note that we have defined L, λ, and W here as averages *over a finite time horizon*. In the formal statement of Theorem 1.1, these quantities are defined as infinite limits.

The assumption that customers depart in the order of arrival is not crucial here. We can make a similar argument when the customers depart out of order. This is illustrated by Figure 1.10. In the figure, the shaded rectangles represent, as before, the time spent in the system by each customer. But now, customers depart out of order. In this example, customer 2 departs first, followed by customer 1, followed by customers 5, 6, 4, and 3. Because the departure process is out of order, $D(t)$ (the dashed line) does not follow the edges of the shaded rectangles.

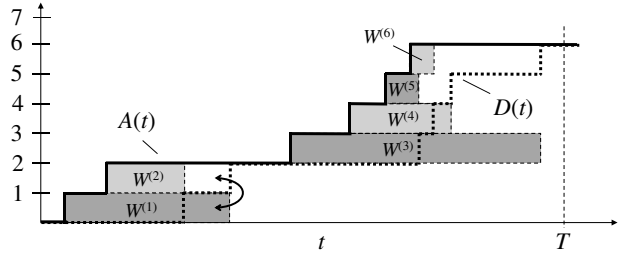


Figure 1.10 Geometric representation of Little's law (customers depart out of order).

Nevertheless, every shaded region that extends outside of the dashed line corresponds precisely to an empty region of equal area that is between $A(t)$ and $D(t)$. For example, the portion of $W^{(1)}$ that extends past $D(t)$ can be mapped to the empty area above it, as shown by the arrow in the figure. In a similar manner, the portions of $W^{(3)}$ and $W^{(4)}$ that extend past $D(t)$ fit exactly into the empty areas above it, though this requires some cutting and rearranging of the rectangles.

In summary, even when customers depart out of order, the total time in the system (i.e., the area of the shaded rectangles, which is $W^{(1)} + \ldots + W^{(N)}$) is exactly equal to the integral of $A(t) - D(t)$ on $[0, T]$. The basic reason this works is the following: *Every unit of time a customer spends waiting in the system contributes exactly one unit to the time integral of the total count of customers in the system.* If the system starts and ends in an empty state, then each customer's time in system is exactly accounted for in the integral on $[0, T]$.

What happens if the system does not end in an empty state? In this case, there would be at least one rectangle $W^{(i)}$ extending past T. So there would be a mismatch between the area of the shaded rectangles and the integral of $A(t) - D(t)$ on $[0, T]$. Nevertheless, it seems reasonable to expect that, over a long time horizon, this mismatch would be small relative to the total integral, and that $L = \lambda W$ would be valid in the limit as $T \to \infty$. This intuition is correct under the assumptions of Theorem 1.1, namely that the limit for the long-term arrival rate λ in (1.1) exists and the limit for the average time in the system W exists.

1.4.2 $H = \lambda G$

It turns out that $L = \lambda W$ is a special case of a more general relation, namely $H = \lambda G$. In this latter formula, G represents the average "cost" or "work" associated with a customer, and H represents the total average cost per time incurred by the system.

More specifically, suppose that customer k arrives to a system at time $A^{(k)}$ and departs for good at time $A^{(k)} + W^{(k)}$. Let $f_k(t)$ denote a weighting function on the time spent in the system by customer k at time t, where $f_k(t) = 0$ for $t \notin [A^{(k)}, A^{(k)} + W^{(k)}]$. The weighting function can be negative, but we require that $\int_0^\infty |f_k(t)|dt = 0$. Define the following quantities:

$$G^{(k)} \equiv \int_{A^{(k)}}^{A^{(k)}+W^{(k)}} f_k(t)\,dt \qquad \text{and} \qquad H(t) \equiv \sum_{k=1}^{\infty} f_k(t).$$

$G^{(k)}$ denotes the total cost or work associated with customer k. $H(t)$ denotes the total cost incurred per time by the system at time t. Analogous to (1.1), define the following limits:

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

Theorem 1.2 *If the limits λ and G exist and are finite and $W^{(k)}/A^{(k)} \to 0$ as $k \to \infty$, then H exists and $H = \lambda G$.*

For a proof see, for example, Wolff (2011). The requirement that $W^{(k)}/A^{(k)} \to 0$ is a technical condition that prevents the departure times from being pushed further and further past the arrival times. Little's law is a special case of this theorem when $f_k(t) = 1$ on the interval $[A^{(k)}, A^{(k)} + W^{(k)}]$.

■ **EXAMPLE 1.6**

A company owns two kinds of machines (type-1 and type-2). Whenever a machine breaks, it is sent to the repair shop. For every hour that a type-i machine is in the shop, the company incurs a cost of c_i, where $c_1 = \$500$ and $c_2 = \$200$. Machines fail at a rate of 1 every 40 hours; half of all failures are type-1 machines are half are type-2. The time to repair a machine is 3 hours, on average, regardless of the type. What is the hourly cost to the company for machine downtime?

Let $f_k(t) = c_{i(k)}$ for $t \in [A^{(k)}, A^{(k)} + W^{(k)}]$ ($f_k(t) = 0$ otherwise), where $A^{(k)}$ is the time of the kth failure, $W^{(k)}$ is the downtime, and $i(k) \in \{1, 2\}$ is the machine type. Then $G^{(k)}$ is the downtime cost of the kth failure. Since each type of failure is equally likely, the average cost per machine failure is $G = 3 \cdot (500 + 200)/2 = \$1,050$. Since $\lambda = 1/40$ per hour, we have $H = (1/40) \cdot \$1,050 = \$26.25$ per hour.

■ **EXAMPLE 1.7**

An amusement park has 5,000 visitors each day. The park is open from 10 am to 10 pm. One of the rides in the park is a roller coaster called the Twisty Twister. Each visitor rides the Twisty Twister, on average, 1.2 times per visit to the park. Throughout the day, the average waiting time for the ride is 30 minutes. What is the average number of people in line at the Twisty Twister?

Let $A^{(k)}$ be the time that visitor k arrives at the amusement park, and let $A^{(k)} + W^{(k)}$ be the visitor's departure time from the park. Let $f_k(t)$ be an indicator function equal to 1 when visitor k is in line for the Twisty Twister at time t ($f_k(t) = 0$ otherwise). Then $G^{(k)}$ is the *total time* that visitor k spends in line at the ride throughout the day. The average over all customers is $G = 1.2 \cdot 0.5 = 0.6$ hours, since each customer takes an average of 1.2 rides and the waiting time for each ride is 0.5 hours. Because the "system" is defined as the amusement park, λ is the arrival rate to the amusement park, not the arrival rate to the ride, so $\lambda = 5,000/12$ per hour. The average number in line at the Twisty Twister is $H = \lambda G = 5,000/12 \cdot 0.6 = 250$. This example illustrates that the weighting function can be used to handle situations where a customer has multiple visits to a subsystem.

1.4.3 Distributional Form of Little's Law

Little's law provides a relationship between the *first* moments of $N(t) \equiv A(t) - D(t)$ (the number of customers in the system at t) and $W^{(k)}$. One might also wonder if it is possible to relate the second moments. The answer is yes. In fact, it is possible to relate all higher moments of $N(t)$ and $W^{(k)}$. Such results come from the *distributional* form of Little's law. While these results relate higher moments, they require a much more restrictive set of assumptions than the main form of Little's law stated in Theorem 1.1.

To state the distributional form of Little's law, consider a system to which customers arrive and from which they depart. The system has the following properties: (1) The arrival process is stationary, (2) customers depart from the system in the order in which they arrive, (3) the time $W^{(k)}$ spent by the kth customer in the system is stationary, and (4) $W^{(k)}$ is independent of the arrival process after the arrival of customer k. Then

$$\Pr\{N(t) \leq j\} = \Pr\{A(W^{(k)}) \leq j\}. \tag{1.4}$$

Equation (1.4) relates the distribution of the number $N(t)$ of customers in the system with the distribution of the number of arrivals $A(\cdot)$ occurring over an interval of length $W^{(k)}$. That is, if one generates a random waiting time $W^{(k)}$ and then generates a random number of arrivals occurring over an interval of length $W^{(k)}$, then this has the same distribution as the number of customers in the system. Various forms of this result are given in Haji and Newell (1971), Brumelle (1972), Keilson and Servi (1988), Bertsimas and Nakazato (1995), and Wolff (2011).

An important special case occurs when the arrival process is Poisson. Then it can be shown that (1.4) implies the following relationship between the jth moments:

$$\mathrm{E}[N(t)(N(t)-1)(N(t)-2)\cdots(N(t)-j+1)] = \lambda^j \mathrm{E}[(W^{(k)})^j]. \qquad (1.5)$$

For example, $j = 2$ gives a relationship between the second moments:

$$\mathrm{E}[N(t)(N(t)-1)] = \lambda^2 \mathrm{E}[(W^{(k)})^2].$$

These equations will also be derived directly for the $M/G/1$ queue; see (6.30) in Section 6.1.5.

In using the distributional form of Little's law, care must be taken to check the assumptions. For example, the requirement that customers depart in the order of arrival does not typically hold for multiserver systems, such as the $M/M/c$ system (though it does hold for the $M/D/c$ queue). This is because customers can pass each other while being served, if there is more than one server. But *for the queue itself*, first-in, first-out (FIFO) is preserved, assuming that customers remain in order in the queue and that no reneging occurs (a customer departing the queue early would violate the FIFO property). Thus, while the distributional law does *not* apply to the full $M/M/c$ system (queue and servers), it does apply to just the queue.

A similar argument can be made for priority systems. FIFO is violated in a priority system because high-priority customers can jump ahead of low-priority customers, so the distributional law does not apply to the system as a whole. But it could apply *separately* to each customer-class queue (provided that the previous assumptions apply). For example, in a two-class $M/G/1$ priority system, the distributional law can be applied to the queue of low-priority customers, and it can be applied separately to the queue of high-priority customers.

Finally, we note that if the arrival process is not renewal (i.e., if the interarrival times are not independent and identically distributed), then the fourth assumption can be violated. Because interarrival times are not independent, there could be a dependency between the arrivals before customer k's arrival and the later arrivals. This means that the waiting time of customer k, which depends on the previous arrivals, could then depend on the subsequent arrivals. Throughout this text, we typically assume that the arrival process is a renewal process.

1.5 General Results

We now present some general results for $G/G/1$ and $G/G/c$ queues, prior to specific model development in later sections. These results will prove useful in many of the following chapters, as well as providing some insight at this early stage.

Table 1.2 summarizes the key notation. Let λ denote the average rate that customers arrive to a queueing system. Let S denote a random service time. The average rate that customers are served (per server) is $\mu \equiv 1/\mathrm{E}[S]$. A measure of total load on the system is the *offered load*, defined as $r \equiv \lambda/\mu = \lambda \mathrm{E}[S]$. Since λ is the average number of customers arriving per unit time and each customer requires an amount

of work $\mathrm{E}[S]$ on average, the offered load $\lambda\mathrm{E}[S]$ represents the amount of work arriving to the system per unit time. Closely related, a measure of traffic congestion is $\rho \equiv \lambda/c\mu$, which is the *traffic intensity* or *utilization*. The traffic intensity is the offered load divided by the number of servers, representing the average amount of work coming to each server per unit time.

Table 1.2 Summary of notation

λ	Average arrival rate
S	Random service time
$\mu \equiv 1/\mathrm{E}[S]$	Average service rate
c	Number of servers
$r \equiv \lambda/\mu$	Offered load
$\rho \equiv \lambda/c\mu$	Traffic intensity or utilization
T, T_q	Random time a customer spends in the system / queue
W, W_q	Average time a customer spends in the system / queue
N, N_q	Random number of customers in the system / queue
L, L_q	Average number of customers in the system / queue

Let T_q represent the random time a customer (in steady state) spends waiting in the queue prior to entering service, and let T represent the random time a customer spends in the system. Then $T = T_q + S$, where S is a random service time. Two often used measures of system performance are the mean waiting time in queue W_q and the mean waiting time in the system W, namely

$$W_q \equiv \mathrm{E}[T_q] \quad \text{and} \quad W \equiv \mathrm{E}[T].$$

Let N_q denote the steady-state number of customers in the queue (a random variable), and let N denote the steady-state number of customers in the system (a random variable). Two measures of interest are the mean number in the queue L_q and the mean number in the system L. Let $p_n = \Pr\{N = n\}$ denote the steady-state probability that there are n customers in the system. Then for a c-server system, L and L_q can be expressed as follows:

$$L \equiv \mathrm{E}[N] = \sum_{n=0}^{\infty} np_n, \qquad L_q \equiv \mathrm{E}[N_q] = \sum_{n=c+1}^{\infty} (n-c)p_n.$$

Using Little's law (Section 1.4), we can establish relationships among the four measures of performance: L, L_q, W, and W_q (Figure 1.11). Specifically, Little's law applied to the system gives $L = \lambda W$ (Example 1.2), and Little's law applied to the queue gives $L_q = \lambda W_q$ (Example 1.3). Also, since $T = T_q + S$ (the time a customer spends in the system is the time spent in the queue plus the time spent in service), taking expectations gives

$$W = W_q + 1/\mu.$$

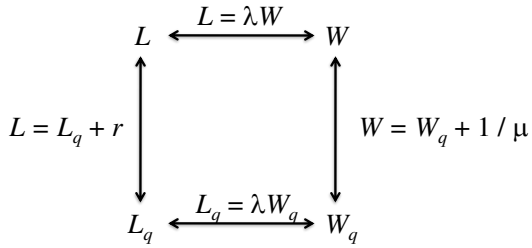


Figure 1.11 Relationships among L, L_q, W, and W_q.

The relation between L and L_q is then obtained from the other three relations:

$$L = \lambda W = \lambda(W_q + 1/\mu) = \lambda W_q + \lambda/\mu = L_q + r. \tag{1.6}$$

In later chapters, we will typically focus on deriving *one* of the four performance measures. The initial derivation may take some effort, but once one of the four measures is obtained, the other three measures follow immediately from the relationships in Figure 1.11.

In the case of a single server, $c = 1$, equation (1.6) has a particular form:

$$r = L - L_q = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} p_n = 1 - p_0.$$

Since $r = \rho$ when $c = 1$, we have

$$\rho = 1 - p_0 \quad \text{or} \quad p_0 = 1 - \rho.$$

That is, for a single-server queue, the fraction of time the system is empty is $1 - \rho$. (This relationship was also obtained directly by applying Little's law to the server in Example 1.4). These results are summarized in Table 1.3.

Table 1.3 Summary of results for $G/G/1$ and $G/G/c$ queues

$L = \lambda W$	$G/G/c$
$L_q = \lambda W_q$	$G/G/c$
$W = W_q + 1/\mu$	$G/G/c$
$L = L_q + r$	$G/G/c$
$r \equiv \lambda/\mu$ = average number of busy servers	$G/G/c$
$\rho \equiv \lambda/c\mu$ = fraction of time server is busy	$G/G/c$
$p_0 = 1 - \rho$	$G/G/1$
$L = L_q + \rho$	$G/G/1$

The offered load r is defined as the ratio of λ and μ. Equation (1.6) shows that r can also be interpreted as *the expected number of customers in service* or

equivalently *the average number of busy servers* (since $r = L - L_q$ and the number in the system minus the number in the queue is the number in service).* The offered load represents a minimum number of servers needed to meet a particular traffic demand. For example, if customers arrive at a rate of $\lambda = 12$ per hour and each customer requires an average service time of $\mathrm{E}[S] = 0.5$ hours, then a minimum of 6 servers is needed to handle the load.

In a similar manner, the traffic intensity $\rho \equiv \lambda/c\mu$ can be interpreted as the fraction of time each server is busy. Since the expected number of busy servers at any instant in steady state is r and there are c available servers, the fraction of time each server is busy is $r/c = \rho$. This assumes symmetry of the servers – that is, there is no inherent preference for any one server to be used more than any other.

It turns out that for steady-state results to exist, we must have $\rho < 1$, or $\lambda < c\mu$. That is, the average rate of arrivals into the system must be strictly less than the maximum average service rate of the system. When $\rho > 1$, customers arrive faster than they can be served, on average, so the queue gets bigger and bigger as time goes on. There is no steady state, since the queue size never settles down. When $\rho = 1$, the arrival rate exactly equals the maximum service rate. In this case, no steady state exists unless arrivals and service times are deterministic and perfectly scheduled (e.g., a queue where customers arrive exactly one minute apart and each customer requires exactly one minute of service). In summary, if one knows the average arrival rate and the average service rate, the minimum number of parallel servers required to guarantee a steady-state solution can be calculated by finding the smallest c such that $\rho = \lambda/c\mu < 1$.

In reality, a queue cannot grow without bound forever. An unbounded queue is a consequence of the modeling assumption that all arriving customers join the queue and remain in the system until served. In reality, when $\lambda > c\mu$ and the queue grows very large, several factors may help stabilize the queue: Customers may choose not to join the queue because it is too long (balking), customers may become impatient and leave the queue after joining (reneging), or customers may be prevented from joining the queue due to space restrictions (blocking). These behaviors are discussed in Section 3.10.

1.6 Simple Bookkeeping for Queues

In this section, we use event-oriented bookkeeping to show how the random events of arrivals and service completions interact to form a queue. Bookkeeping has to do with updating the system status whenever events occur, recording items of interest, and calculating measures of effectiveness. *Event-oriented* bookkeeping updates the system state *only when events occur* (e.g., when customers arrive or depart). The master clock is increased by a possibly different amount each time. (This is in

*This interpretation is valid for a $G/G/c$ queue. For a queue with blocking, such as an $M/M/c/K$ queue, the offered load can be interpreted as the expected number of customers in service *in a hypothetical scenario in which the system has an infinite number of servers*.

contrast to *time-oriented* bookkeeping in which the master clock is increased by a fixed amount each step, regardless of when events occur.)

The event-oriented approach is illustrated by an example using the arrival and service data given in Table 1.4. Such data might be collected by recording the times when customers arrive to a queueing system as well as the starting and ending times of service for each customer. From this data, we seek to establish how the queue forms in time. The analysis is obtained under the assumption of *a single server with FCFS discipline*.

Table 1.4 Input data

n	1	2	3	4	5	6	7	8	9	10	11	12
Arrival time of cust. n	0	2	3	6	7	8	12	14	19	20	24	26
Service time of cust. n	1	3	6	2	1	1	4	2	5	1	1	3

To conduct this analysis, we define a number of variables associated with each customer, as shown in Table 1.5. The first two variables, $A^{(n)}$ and $S^{(n)}$, are inputs, and the remaining variables can be derived from these inputs. Various relationships among the variables are given in the table. For example, the time that a customer departs the system is the time the customer begins service plus the customer's service time. A key relationship is $U^{(n+1)} = \max\{D^{(n)}, A^{(n+1)}\}$, which says that customer $n+1$ begins service when customer n departs; however, customer $n+1$ cannot begin service prior to his or her own arrival, which is the reason for the maximum of the two variables. This relationship is specific to a single-server FCFS queue, while the other relationships in the table hold in more general situations.

To find the queue waiting times, we observe that $W_q^{(n)}$ and $W_q^{(n+1)}$ of two successive customers in *any* FCFS single-server queue (deterministic or otherwise)

Table 1.5 Notation and basic relationships

Variable	Definition	Sample Relationship
$A^{(n)}$	Arrival time of cust. n	
$S^{(n)}$	Service time of cust. n	
$T^{(n)}$	Interarrival time cust. n and $n+1$	$T^{(n)} = A^{(n+1)} - A^{(n)}$
$U^{(n)}$	Time cust. n starts service	$U^{(n+1)} = \max\{D^{(n)}, A^{(n+1)}\}$
$D^{(n)}$	Departure time of cust. n	$D^{(n)} = U^{(n)} + S^{(n)}$
$W_q^{(n)}$	Time in queue of cust. n	$W_q^{(n)} = U^{(n)} - A^{(n)}$
$W^{(n)}$	Time in system of cust. n	$W^{(n)} = W_q^{(n)} + S^{(n)}$

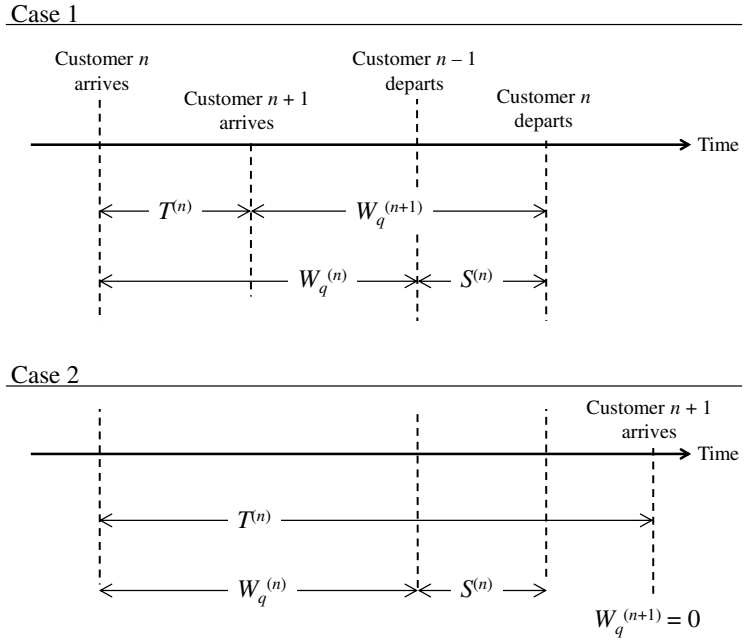


Figure 1.12 Lindley's equation: Successive $G/G/1$ waiting times.

are related by the simple recurrence relation

$$\boxed{W_q^{(n+1)} = \max\{W_q^{(n)} + S^{(n)} - T^{(n)}, 0\}.} \tag{1.7}$$

This is called *Lindley's equation* and is an important general relation that is utilized in later portions of the text. The equation can be seen via the diagrams in Figure 1.12. The equation says that the wait in queue of an arriving customer is the wait in queue of the previously arriving customer plus that customer's service time, minus the interarrival time between the two customers (Case 1 in the figure). However, with a long interarrival time, this value could be negative, so the maximum in (1.7) insures that $W_q^{(n+1)}$ is never negative, which is illustrated by Case 2.

Lindley's equation can also be obtained using the relationships from Table 1.5:

$$\begin{aligned} W_q^{(n+1)} &= U^{(n+1)} - A^{(n+1)} \\ &= \max\{D^{(n)}, A^{(n+1)}\} - A^{(n+1)} \\ &= \max\{D^{(n)} - A^{(n+1)}, 0\} \\ &= \max\{U^{(n)} + S^{(n)} - A^{(n)} - T^{(n)}, 0\} \\ &= \max\{W_q^{(n)} + S^{(n)} - T^{(n)}, 0\}. \end{aligned}$$

Table 1.6 shows bookkeeping results for the input data, based on the relationships in Table 1.5. These values are easily obtained in a spreadsheet by entering a formula

for each variable and copying the formula down each column. Note that $W_q^{(n)}$ can be obtained just from the columns for $S^{(n)}$ and $T^{(n)}$ via Lindley's equation (1.7), so it may not be necessary to track all of the variables in a bookkeeping approach.

Table 1.6 Event-based bookkeeping

Customer	$A^{(n)}$	$S^{(n)}$	$T^{(n)}$	$U^{(n)}$	$D^{(n)}$	$W_q^{(n)}$	$W^{(n)}$
1	0	1	2	0	1	0	1
2	2	3	1	2	5	0	3
3	3	6	3	5	11	2	8
4	6	2	1	11	13	5	7
5	7	1	1	13	14	6	7
6	8	1	4	14	15	6	7
7	12	1	2	15	16	3	4
8	14	2	5	16	18	2	4
9	19	5	1	19	24	0	5
10	20	1	4	24	25	4	5
11	24	1	2	25	26	1	2
12	26	3	–	26	29	0	3

To compute measures of effectiveness, the sample averages for W_q and W are the averages of the columns for $W_q^{(n)}$ and $W^{(n)}$, that is, $W_q = 29/12$ and $W = 56/12$. To determine L and L_q, we must first define the time horizon over which the sample averages are computed. Since the last departure occurs at time 29, a natural time horizon is $[0, 29]$. Over this interval, the system starts and ends in an empty state, so Little's law provides an exact relationship of the sample values for L, λ, and W (see Figure 1.9 and the associated discussion). The sample arrival rate over this time interval is $\lambda = 12/29$. Thus,

$$L_q = \lambda W_q = \frac{12}{29} \cdot \frac{29}{12} = 1, \quad \text{and} \quad L = \lambda W = \frac{12}{29} \cdot \frac{56}{12} = \frac{56}{29}.$$

Alternatively, we can determine L directly from the time average of $N(t)$, which is the number of customers in the system at time t. Assuming that the system starts in an empty state (just prior to $t = 0$), we write

$$N(t) = \{\text{number of arrivals in } [0, t]\} - \{\text{number of departures in } [0, t]\}. \quad (1.8)$$

Figure 1.13 shows the resulting sample path of $N(t)$. At every arrival point, $N(t)$ increases by 1, at every departure point, it decreases by 1. The time average is

$$L = \frac{1}{29} \int_0^{29} N(t)\, dt = \frac{1}{29}(1 \cdot 10 + 2 \cdot 9 + 3 \cdot 4 + 4 \cdot 4) = \frac{56}{29}.$$

(The time average is obtained by observing that $N(t) = 1$ for 10 time units, $N(t) = 2$ for 9 time units, and so forth.) The sample average for L_q can be determined in a similar manner as the time average of $N_q(t)$, where $N_q(t) = \max\{N(t) - 1, 0\}$.

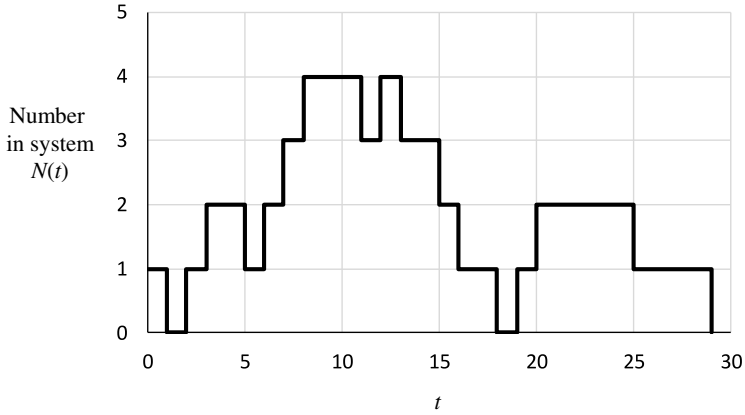


Figure 1.13 Sample path for queueing process.

Note that the bookkeeping approach is based on the sample-path observations of the data. No assumptions are required on the probability laws for the interarrival times or service times, since the results are derived directly from the data, regardless of the probability laws that generated them.

1.7 Introduction to the QtsPlus Software

Today, spreadsheets are an indispensable tool for engineers and operations research specialists. Several papers have discussed the application of spreadsheets in the various operations research disciplines, such as optimization and queueing theory (Bodily, 1986; Leon et al., 1996; Grossman, 1999). To facilitate learning, a collection of spreadsheet queueing models, collectively known as *QtsPlus*, is available with this textbook. Most of the models analyzed in this textbook are implemented as spreadsheet models in *QtsPlus*. See Appendix E for instructions to install and run the software.

We illustrate how to use QtsPlus with an example involving the stationary distribution of a Markov chain (see Example 2.6 from the next chapter). Follow the instructions in Appendix E to start the software. Once it is active, select the **Basic** model category from the list provided, then select the **Discrete-Time Markov Chain** model from the available list. Once the model workbook (marchain.xlsm) is open, enter 2 into the input field: **Number of States**. A pop-up message box may appear asking,

> This will cause existing model parameters to be discarded. Do you wish to continue?

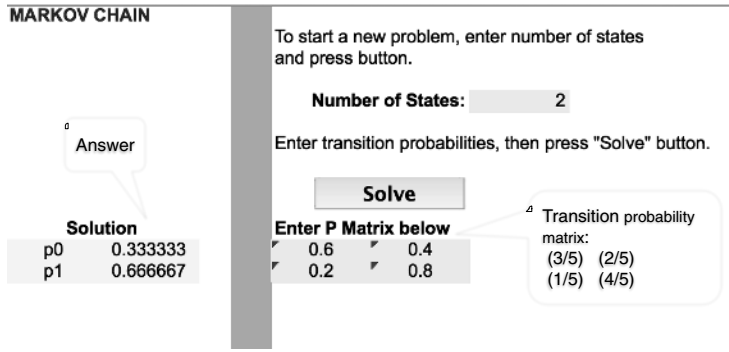


Figure 1.14 QtsPlus solution to Example 2.6.

Press the **Yes** button to set up a new $\boldsymbol{P}$ matrix. Using Excel formulas, fill the respective cells of the $\boldsymbol{P}$ matrix in the worksheet with the initial parameters as shown below.

$$\begin{array}{ll} = 3/5 & = 2/5 \\ = 1/5 & = 4/5 \end{array}$$

Press the **Solve** button. The answer appears on the left side of the worksheet (Figure 1.14) and coincides with the stationary solution $\pi = (\frac{1}{3}, \frac{2}{3})$ obtained in Example 2.6.

PROBLEMS

1.1. Discuss the following queueing situations in terms of the characteristics given in Section 1.2.

- **(a)** Aircraft landing at an airport.
- **(b)** Supermarket checkout procedures.
- **(c)** Post-office or bank customer windows.
- **(d)** Toll booths on a bridge or highway.
- **(e)** Gasoline station with several pump islands.
- **(f)** Automatic car wash facility.
- **(g)** Telephone calls coming into a customer information system.
- **(h)** Appointment patients coming into a doctor's office.
- **(i)** Tourists wishing a guided tour of the White House.
- **(j)** Electronic components on an assembly line consisting of three operations and one inspection at end of line.
- **(k)** Processing of programs coming from a number of independent sources on a local area network into a central computer.

1.2. Give three examples of a queueing situation other than those listed in Problem 1.1, and discuss them in terms of the basic characteristics of Section 1.2.

1.3. The Carry Out Curry House, a fast-food Indian restaurant, must decide on how many parallel service channels to provide. They estimate that, during

the rush hours, the average number of arrivals per hour will be approximately 40. They also estimate that, on average, a server will take about 5.5 min to serve a typical customer. Using only this information, how many service channels will you recommend they install?

1.4. Fluffy Air, a small local feeder airline, has a customer-service call center. They want to know how many slots to provide for telephone callers to be placed on hold. They plan to have enough service representatives so that the average waiting time on hold for a caller will be 75 seconds or less during the busiest period of the day. They estimate the average call-in rate to be 3 per minute during this time. What would you advise?

1.5. The Outfront BBQ Rib Haven does carry out only. During peak periods, two servers are on duty. The owner notices that during these periods, the servers are almost never idle. She estimates the percent idle time of each server to be 1 percent. Ideally, the percent idle time would be 10 percent to allow time for important breaks.

(a) If the owner decides to add a third server during these times, how much idle time would each server have then?

(b) Suppose that by adding the third server, the pressure on the servers is reduced, so they can work more carefully, but their service output rate is reduced by 20 percent. What now is the percent time each would be idle?

(c) Suppose, instead, that the owner decides to hire an aid (at a much lower salary) who servers as a gofer for the two servers, rather than hiring another full server. This allows the two servers to decrease their average service *time* by 20 percent (relative to the original service times). What now is the percent idle time of each of the two servers?

1.6. The Happidaiz frozen yogurt stand does a thriving business on warm summer evenings. Even so, there is only a single person on duty at all times. It is known that the service time (dishing out the yogurt and collecting the money) is normally distributed with mean 2.5 min and standard deviation 0.5 min. (Although the normal distribution allows for negative values, the standard deviation with respect to the mean is small so that negative values are more than 4 standard deviations below the mean and the probability of negative values is essentially zero.) You arrive on a particular evening to get your favorite crunchy chocolate yogurt cone and find 8 people ahead of you. Estimate the average time until you get the first lick. What is the probability that you will have to wait more than 0.5 h? [*Hint*: Remember that the sum of normal random variables is itself normally distributed.]

1.7. A certain football league consists of 32 teams. Each team has 67 active players. There is a draft each year for teams to acquire new players. Each team acquires 7 new players per year in the draft. The number of active players on each team must always be 67. Thus, each team must cut some existing players each year to make room for the new players.

(a) Assuming that a football player can only join a team by being selected in the draft, estimate the average career length of a football player in the league.

(b) Now, suppose that a player can join a team in one of two ways: (1) by being selected in the draft, as before, or (2) by signing directly with a team outside the draft. Suppose further that the average career length of a football player is known to be 3.5 years. Under the same assumptions as before, estimate the average number of players who enter the league each year *without being drafted*.

1.8. The following table gives enrollment statistics for undergraduates at a university. From this data, estimate the average length of time that an undergraduate is enrolled at the university (this average should include students who graduate as well as students who transfer or drop out).

Year	New Students		Total
	First Year Students	Transfer Students	Enrollment
1	1,820	2,050	16,800
2	1,950	2,280	16,700
3	1,770	2,220	17,100
4	1,860	2,140	16,400
5	1,920	2,250	17,000

1.9. You are selling your home. You observe that at any given time there are typically about 50 homes for sale in your area. New homes go on the market at a rate of about 5 per week. About how long will it take to sell your home? What assumptions are made to arrive at your answer?

1.10. Suppose that an $M/G/1/K$ queue has a blocking probability of $p_k = 0.1$ with $\lambda = \mu = 1$ and $L = 5$. Find W, W_q, and p_0.

1.11. Suppose that it costs \$3 to make one dose of the small pox vaccine. Once a dose is made, its shelf life is 90 days, after which it can no longer be used. It is desired to have, on average, 300 million doses available at any given time.

(a) What is the yearly cost to implement this plan?

(b) Suppose now that the shelf life of a vaccine is randomly distributed according to an Erlang distribution with a mean of 90 days and a standard deviation of 30 days. What is the yearly cost to implement this plan?

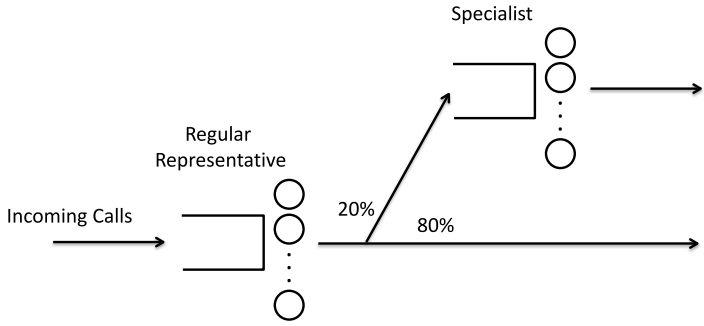
(c) Suppose that a vaccine with a longer shelf life can be made, but at a greater cost. It is found that the cost to produce a vaccine with a shelf life of x days is equal to $a + bx^2$, where $a = \$2.50$ and $b = \$0.00005$. What is the shelf life that minimizes the yearly cost?

1.12. Customers who have purchased a Delta laptop may call a customer support center to get technical help. Initially, a call is handled by a regular service

representative. If the problem cannot be handled by a regular service representative, the call is transferred to a specialist. Twenty percent of all calls are transferred to a specialist. On average, there are 40 customers being served or waiting to be served by a regular representative. On average, there are 10 customers being served or waiting to be served by a specialist. The average rate of incoming calls is 100 per hour. There are 30 regular representatives and 10 specialists.

(a) What is the average time spent in the system for an arbitrary customer? State any assumptions you make to answer this question.

(b) What is the average time spent in the system for a customer who needs to talk to a specialist?



1.13. Consider the following (very) simplified model of Social Security. Every year, 3 million people turn 65. A person begins to receive Social Security benefits when s/he reaches an age of 65 years. An individual (over the age of 65) has a 5% chance of dying each year, independent of all else. Social Security benefits are $40,000 per person per year.

(a) On average, how long does a person receive Social Security benefits?

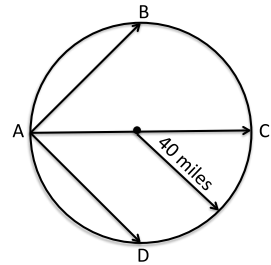
(b) What is the average total yearly payout in Social Security benefits?

1.14. Planes arrive at a circular sector of airspace according to a Poisson process with rate 20 arrivals per hour. The radius of the sector is 40 miles. Each plane travels at a speed of 400 miles per hour. There are 4 possible entrance / exit points in the sector, as shown. An aircraft is equally like to arrive and depart from any of points A, B, C, and D (but an aircraft cannot enter and exit from the same point). For example, the probability that an aircraft arrives at point A is 1/4. Given that an aircraft arrives at A, the probability that it exits at B, C, or D is 1/3 each. Assume that aircraft flights are straight paths and there are no collisions or conflict avoidance maneuvers in the sector.

(a) What is the average path length across the sector?

(b) What is the average number of aircraft in the sector?

(c) If we suppose that aircraft sometimes execute avoidance maneuvers to prevent conflicts/collisions, would the answer in (b) go up or down?

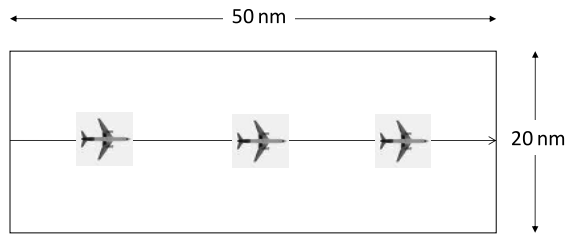


1.15. The length of time that a person owns a car before buying a new one has an Erlang-3 distribution with a mean of 5 years. Suppose that there are approximately 150 million cars in the United States.

(a) Assuming that a person's old car is destroyed when he or she buys a new car, how many cars does the auto industry expect to sell each year?

(b) Now assume that a person's old car is sold to somebody else when that person buys a new car. The person who buys the used car keeps it for period of time following an Erlang-3 distribution with a mean of 7 years. When that person buys another used car, his or her previous used car is assumed to be destroyed. Under the same previous assumptions, how many new cars does the auto industry expect to sell each year?

1.16. Aircraft enter a sector as shown in the following figure. The sector length is 50 nautical miles (nm). The spacing between aircraft as they enter the sector is 5 nm plus an exponentially distributed random variable with a mean of 1 nm. Suppose that aircraft travel at 400 knots (nautical miles per hour). What is the average number of aircraft in a sector?



1.17. Table 1.7 gives observations regarding customers at a single-server FCFS queue.

(a) Compute the average time in the queue and the average time in the system.

(b) Calculate the average system waiting time of those customers who had to wait for service (i.e., exclude those who were immediately taken into service). Calculate the average length of the queue, the average number in the system, and the fraction of idle time of the server.

Table 1.7 Data for Problem 1.17

Customer	Interarrival Time	Service Time
1	1	3
2	9	7
3	6	9
4	4	9
5	7	10
6	9	4
7	5	8
8	8	5
9	4	5
10	10	3
11	6	6
12	12	3
13	6	5
14	8	4
15	9	9
16	5	9
17	7	8
18	8	6
19	8	8
20	7	3

1.18. Items arrive at an initially unoccupied inspection station at a uniform rate of one every 5 min. With the time of the first arrival set equal to 5, the chronological times for inspection completion of the first 10 items were observed to be 7, 17, 23, 29, 35, 38, 39, 44, 46, and 60, respectively. By manual simulation of the operation for 60 min, using these data, develop sample results for the mean number in system and the percentage idle time experienced.

1.19. Table 1.8 lists the arrival times and service durations for customers in a FCFS single-server queue. From this data, compute L_q (the time-average number in queue) and $L_q^{(A)}$ (the average number in queue as seen by arriving customers). For L_q, use a time horizon of [0, 15.27], where 15.27 is the time that the last customer exits the system. Assume the system is empty at $t = 0$.

Table 1.8 Data for Problem 1.19

Arrival Time (min)	Service Duration (min)
1	2.22
2	1.76
3	2.13
4	0.14
5	0.76
6	0.70
7	0.47
8	0.22
9	0.18
10	2.41
11	0.41
12	0.46
13	1.37
14	0.27
15	0.27