Queueing Models

Simulation is often used in the analysis of queueing models. In a simple but typical queueing model, shown in Figure 1, customers arrive from time to time and join a *queue* (waiting line), are eventually served, and finally leave the system. The term *customer* refers to any type of entity that can be viewed as requesting service from a system. Therefore, many service facilities, production systems, repair and maintenance facilities, communications and computer systems, and transport and material-handling systems can be viewed as queueing systems.

Queueing models, whether solved mathematically or analyzed through simulation, provide the analyst with a powerful tool for designing and evaluating the performance of queueing systems. Typical measures of system performance include server utilization (percentage of time a server is busy), length of waiting lines, and delays of customers. Quite often, when designing or attempting to improve a queueing system, the analyst (or decision maker) is involved in tradeoffs between server utilization and customer satisfaction in terms of line lengths and delays. Queueing theory and simulation analysis are used to predict these measures of system performance as a function of the input parameters. The input parameters include the arrival rate of customers, the service demands of customers, the rate at which a server works, and the number and arrangement of servers. To a certain degree, some of the input parameters are under management's direct control. Consequently, the performance measures could be under their indirect control, provided that the relationship between the performance measures and the input parameters is adequately understood for the given system.

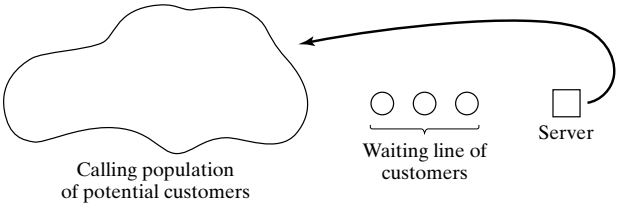

Figure 1 Simple queueing model.

For relatively simple systems, these performance measures can be computed mathematically—at great savings in time and expense as compared with the use of a simulation model—but, for realistic models of complex systems, simulation is usually required. Nevertheless, analytically tractable models, although usually requiring many simplifying assumptions, are valuable for rough-cut estimates of system performance. These rough-cut estimates may then be refined by use of a detailed and more realistic simulation model while also providing a way to verify that the simulation model has been programmed correctly. Simple models are also useful for developing an understanding of the dynamic behavior of queueing systems and the relationships between various performance measures. This chapter will not develop the mathematical theory of queues but instead will discuss some of the well-known models. For an elementary treatment of queueing theory, the reader is referred to the survey chapters in Hillier and Lieberman [2005] or Winston [2004]. More extensive treatments with a view toward applications are given by Cooper [1990], Gross and Harris [1997], Hall [1991] and Nelson [1995]. The latter two texts especially emphasize engineering and management applications.

This chapter discusses the general characteristics of queues, the meanings and relationships of the important performance measures, estimation of the mean measures of performance from a simulation, the effect of varying the input parameters, and the mathematical solution of a small number of important and basic queueing models.

1 Characteristics of Queueing Systems

The key elements of a queueing system are the customers and servers. The term *customer* can refer to people, machines, trucks, mechanics, patients, pallets, airplanes, e-mail, cases, orders, or dirty clothes—anything that arrives at a facility and requires service. The term *server* might refer to receptionists, repair personnel, mechanics, medical personnel, automatic storage and retrieval machines (e.g., cranes), runways at an airport, automatic packers, order pickers, CPUs in a computer, or washing machines—any resource (person, machine, etc.) that provides the requested service. Although the terminology employed will be that of a customer arriving at a service facility, sometimes the server moves to the customer; for example, a repair person moving to a broken machine. This in no way invalidates the models but is merely a matter of terminology. Table 1 lists a number of different systems together with a subsystem consisting of arriving customers and one or more servers. The remainder of this section describes the elements of a queueing system in more detail.

Table 1 Examples of Queueing Systems

System	*Customers*	*Server(s)*
Reception desk	People	Receptionist
Repair facility	Machines	Repair person
Garage	Trucks	Mechanic
Airport security	Passengers	Baggage x-ray
Hospital	Patients	Nurses
Warehouse	Pallets	Fork-lift Truck
Airport	Airplanes	Runway
Production line	Cases	Case-packer
Warehouse	Orders	Order-picker
Road network	Cars	Traffic light
Grocery	Shoppers	Checkout station
Laundry	Dirty linen	Washing machines/dryers
Job shop	Jobs	Machines/workers
Lumberyard	Trucks	Overhead crane
Sawmill	Logs	Saws
Computer	Email	CPU, disk
Telephone	Calls	Exchange
Ticket office	Football fans	Clerk
Mass transit	Riders	Buses, trains

1.1 The Calling Population

The population of potential customers, referred to as the *calling population*, may be assumed to be finite or infinite. For example, consider the personal computers of the employees of a small company that are supported by an IT staff of three technicians. When a computer fails, needs new software, etc., it is attended by one of the IT staff. The computers are customers who arrive at the instant they need attention. The IT staff are the servers who provide repairs, software updates, etc. The calling population is finite and consists of the personal computers at the company.

In systems with a large population of potential customers, the calling population is usually assumed to be infinite. For such systems, this assumption is usually innocuous and, furthermore, it might simplify the model. Examples of infinite populations include the potential customers of a restaurant, bank, or other similar service facility and also the personal computers of the employees of a very large company. Even though the actual population could be finite but large, it is generally safe to use infinite population models, provided that the number of customers being served or waiting for service at any given time is a small proportion of the population of potential customers.

The main difference between finite and infinite population models is how the arrival rate is defined. In an infinite population model, the arrival rate (i.e., the average number of arrivals per unit of time) is not affected by the number of customers who have left the calling population and joined the queueing system. When the arrival process is homogeneous over time (e.g., there are no rush hours),

the arrival rate is usually assumed to be constant. On the other hand, for finite calling-population models, the arrival rate to the queueing system does depend on the number of customers being served and waiting. To take an extreme case, suppose that the calling population has one member, for example, a corporate jet. When the corporate jet is being serviced by the team of mechanics who are on duty 24 hours per day, the arrival rate is zero, because there are no other potential customers (jets) who can arrive at the service facility (team of mechanics). A more typical example is five hospital patients assigned to a single nurse. When all patients are resting and the nurse is idle, the arrival rate is at its maximum since any of the patients could call the nurse for assistance in the next instant. At those times when all five patients have called for the nurse (four are waiting for the nurse and one is being served) the arrival rate is zero; that is, no arrival is possible until the nurse finishes with a patient, in which case the patient returns to the calling population and becomes a potential arrival. It may seem odd that the arrival rate is at its maximum when all five patients are resting. But the arrival rate is defined as the expected number of arrivals in the next unit of time, so this expectation is largest when all patients could potentially call in the next unit of time.

1.2 System Capacity

In many queueing systems, there is a limit to the number of customers that may be in the waiting line or system. For example, an automatic car wash might have room for only 10 cars to wait in line to enter the mechanism. It might be too dangerous (or illegal) for cars to wait in the street. An arriving customer who finds the system full does not enter but returns immediately to the calling population. Some systems, such as in-person concert ticket sales for students, may be considered as having unlimited capacity, since there are no limits on the number of students allowed to wait to purchase tickets. As will be seen later, when a system has limited capacity, a distinction is made between the arrival rate (i.e., the number of arrivals per time unit) and the effective arrival rate (i.e., the number who arrive and enter the system per time unit).

1.3 The Arrival Process

The arrival process for infinite-population models is usually characterized in terms of interarrival times of successive customers. Arrivals may occur at scheduled times or at random times. When at random times, the interarrival times are usually characterized by a probability distribution. In addition, customers may arrive one at a time or in batches. The batch may be of constant size or of random size.

The most important model for random arrivals is the Poisson arrival process. If A_n represents the interarrival time between customer $n-1$ and customer n (A_1 is the actual arrival time of the first customer), then, for a Poisson arrival process, A_n is exponentially distributed with mean $1/\lambda$ time units. The arrival rate is λ customers per time unit. The number of arrivals in a time interval of length t, say $N(t)$, has the Poisson distribution with mean λt customers.

The Poisson arrival process has been employed successfully as a model of the arrival of people to restaurants, drive-in banks, and other service facilities; the arrival of telephone calls to a call center; the arrival of demands, or orders for a service or product; and the arrival of failed components or

machines to a repair facility. Typically, the Poisson arrival process is used to describe a large calling population from which customers make independent decisions about when to arrive.

A second important class of arrivals is scheduled arrivals, such as patients to a physician's office or scheduled airline flight arrivals to an airport. In this case it is usually easier to model the positive or negative deviations from the scheduled arrival time, rather than the interarrival times.

A third situation occurs when at least one customer is assumed to always be present in the queue, so that the server is never idle because of a lack of customers. For example, the customers may represent raw material for a product, and sufficient raw material is assumed to be always available.

For finite-population models, the arrival process is characterized in a completely different fashion. Define a customer as *pending* when that customer is outside the queueing system and a member of the potential calling population. For example, a hospital patient is *pending* when they are resting, and becomes *not pending* the instant they call for the nurse. Define a *runtime* of a given customer as the length of time from departure from the queueing system until that customer's next arrival to the queue. Let $A_1^{(i)}, A_2^{(i)}, \ldots$ be the successive runtimes of customer i; let $S_1^{(i)}, S_2^{(i)}, \ldots$ be the corresponding successive service times; and let $W_{Q1}^{(i)}, W_{Q2}^{(i)}, \ldots$ be the corresponding waiting times for service to begin on each visit to the queueing system. Thus, $W_n^{(i)} = W_{Qn}^{(i)} + S_n^{(i)}$ is the total time spent in system by customer i during the nth visit. Figure 2 illustrates these concepts for patient 3 in the hospital example. The total arrival process is the superposition of the arrival times of all customers. Figure 2 shows the first and second arrival of patient 3, but these two times are not necessarily two successive arrivals to the system. For instance, if it is assumed that all patients are pending at time 0, the first arrival to the system occurs at time $A_1 = \min\{A_1^{(1)}, A_1^{(2)}, A_1^{(3)}, A_1^{(4)}, A_1^{(5)}\}$. If $A_1 = A_1^{(2)}$, then patient 2 is the first arrival (i.e., the first to call for the nurse) after time 0. As discussed earlier, the arrival rate is not constant but is a function of the number of pending customers.

One important application of finite-population models is the machine-repair problem. The machines are the customers, and a runtime is also called *time to failure*. When a machine fails, it *arrives* at the queueing system (the repair facility) and remains there until it is *served* (repaired). Times to failure for a given class of machine have been characterized by the exponential, the Weibull, and the gamma distributions. Models with an exponential runtime are sometimes analytically tractable; an example is given in Section 5. Successive times to failure are usually assumed to be statistically

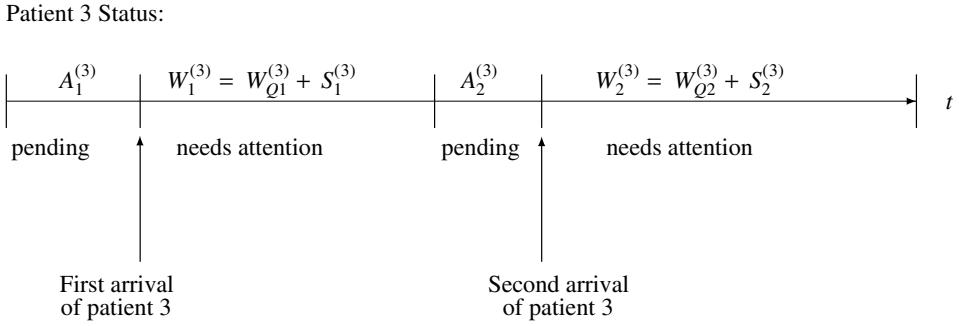

Figure 2 Arrival process for a finite-population model.

independent, but they could depend on other factors, such as the age of a machine since its last major overhaul.

1.4 Queue Behavior and Queue Discipline

Queue behavior refers to the actions of customers while in a queue waiting for service to begin. In some situations, there is a possibility that incoming customers will balk (leave when they see that the line is too long), renege (leave after being in the line when they see that the line is moving too slowly), or jockey (move from one line to another if they think they have chosen a slow line).

Queue discipline refers to the logical ordering of customers in a queue and determines which customer will be chosen for service when a server becomes free. Common queue disciplines include first-in-first-out (FIFO), last-in-first-out (LIFO), service in random order (SIRO), shortest processing time first (SPT), and service according to priority (PR). In a manufacturing system, queue disciplines are sometimes based on due dates and on expected processing time for a given type of job. Notice that a FIFO queue discipline implies that services begin in the same order as arrivals, but that customers could leave the system in a different order because of different-length service times.

1.5 Service Times and the Service Mechanism

The service times of successive arrivals are denoted by $S_1, S_2, S_3, \ldots$ They may be constant or of random duration. In the latter case, $\{S_1, S_2, S_3, \ldots\}$ is usually characterized as a sequence of independent and identically distributed random variables. The exponential, Weibull, gamma, lognormal, and truncated normal distributions have all been used successfully as models of service times in different situations. Sometimes services are identically distributed for all customers of a given type or class or priority, whereas customers of different types might have completely different service-time distributions. In addition, in some systems, service times depend upon the time of day or upon the length of the waiting line. For example, servers might work faster than usual when the waiting line is long, thus effectively reducing the service times.

A queueing system consists of a number of service centers and interconnecting queues. Each service center consists of some number of servers, c, working in parallel; that is, upon getting to the head of the line, a customer takes the first available server. Parallel service mechanisms are either single server ($c = 1$), multiple server ($1 < c < \infty$), or unlimited servers ($c = \infty$). A self-service facility is usually characterized as having an unlimited number of servers.

Example 1

Consider a discount warehouse where customers may either serve themselves or wait for one of three clerks, then finally leave after paying a single cashier. The system is represented by the flow diagram in Figure 3. The subsystem, consisting of queue 2 and service center 2, is shown in more detail in Figure 4. Other variations of service mechanisms include batch service (a server serving several customers simultaneously) and a customer requiring several servers simultaneously. In the discount warehouse, a clerk might pick several small orders at the same time, but it may take two of the clerks to handle one heavy item.

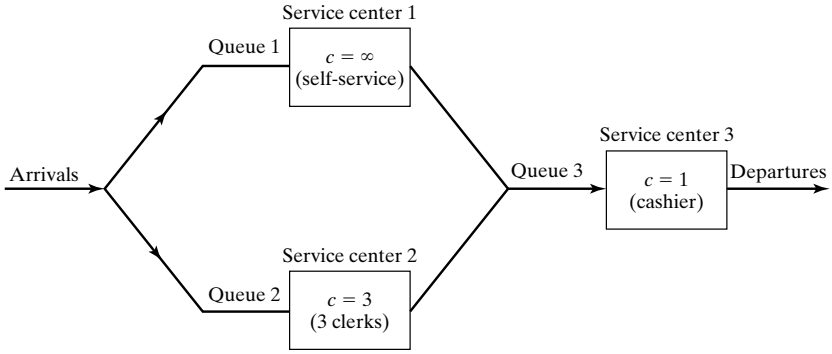

Figure 3 Discount warehouse with three service centers.

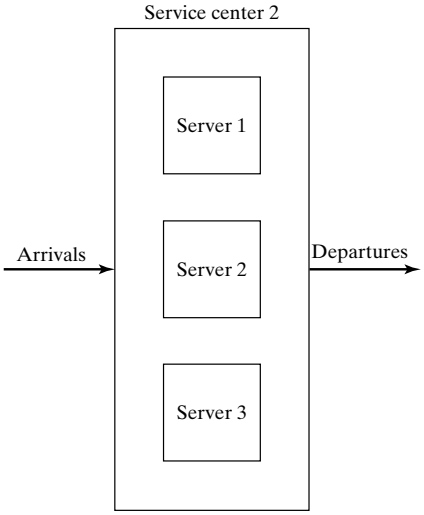

Figure 4 Service center 2, with $c = 3$ parallel servers.

Example 2

A candy manufacturer has a production line that consists of three machines separated by inventory-in-process buffers. The first machine makes and wraps the individual pieces of candy, the second packs 50 pieces in a box, and the third machine seals and wraps the box. The two inventory buffers have capacities of 1000 boxes each. As illustrated by Figure 5, the system is modeled as having three service centers, each center having $c = 1$ server (a machine), with queue capacity constraints between machines. It is assumed that a sufficient supply of raw material is always available at the first queue. Because of the queue capacity constraints, machine 1 shuts down whenever its inventory

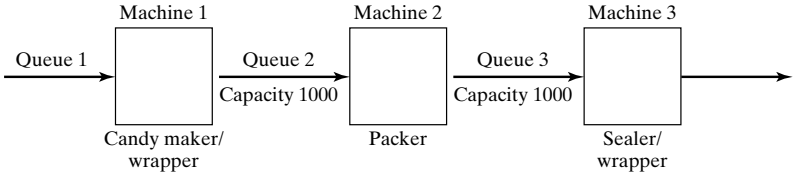

Figure 5 Candy-production line.

buffer (queue 2) fills to capacity, and machine 2 shuts down whenever its buffer empties. In brief, the system consists of three single-server queues in series with queue capacity constraints and a continuous arrival stream at the first queue.

2 Queueing Notation

Recognizing the diversity of queueing systems, Kendall [1953] proposed a notational system for parallel server systems which has been widely adopted. An abridged version of this convention is based on the format $A/B/c/N/K$. These letters represent the following system characteristics:

A represents the interarrival-time distribution.

B represents the service-time distribution.

c represents the number of parallel servers.

N represents the system capacity.

K represents the size of the calling population.

Common symbols for A and B include M (exponential or Markov), D (constant or deterministic), E_k (Erlang of order k), PH (phase-type), H (hyperexponential), G (arbitrary or general), and GI (general independent).

For example, $M/M/1/\infty/\infty$ indicates a single-server system that has unlimited queue capacity and an infinite population of potential arrivals. The interarrival times and service times are exponentially distributed. When N and K are infinite, they may be dropped from the notation. For example, $M/M/1/\infty/\infty$ is often shortened to $M/M/1$. The nurse attending 5 hospital patients might be represented by $M/M/1/5/5$.

Additional notation used throughout the remainder of this chapter for parallel server systems is listed in Table 2. The meanings may vary slightly from system to system. All systems will be assumed to have a FIFO queue discipline.

Table 2 Queueing Notation for Parallel Server Systems

P_n	Steady-state probability of having n customers in system
$P_n(t)$	Probability of n customers in system at time t
λ	Arrival rate
λ_e	Effective arrival rate
μ	Service rate of one server
ρ	Server utilization
A_n	Interarrival time between customers $n-1$ and n
S_n	Service time of the nth arriving customer
W_n	Total time spent in system by the nth arriving customer
W_n^Q	Total time spent waiting in queue by customer n
$L(t)$	The number of customers in system at time t
$L_Q(t)$	The number of customers in queue at time t
L	Long-run time-average number of customers in system
L_Q	Long-run time-average number of customers in queue
w	Long-run average time spent in system per customer
w_Q	Long-run average time spent in queue per customer

3 Long-Run Measures of Performance of Queueing Systems

The primary long-run measures of performance of queueing systems are the long-run time-average number of customers in the system (L) and in the queue (L_Q), the long-run average time spent in system (w) and in the queue (w_Q) per customer, and the server utilization, or proportion of time that a server is busy (ρ). The term *system* usually refers to the waiting line plus the service mechanism, but, in general, can refer to any subsystem of the queueing system; on the other hand, the term *queue* refers to the waiting line alone. Other measures of performance of interest include the long-run proportion of customers who are delayed in queue longer than t_0 time units, the long-run proportion of customers turned away because of capacity constraints, and the long-run proportion of time the waiting line contains more than k_0 customers.

This section defines the major measures of performance for a general $G/G/c/N/K$ queueing system, discusses their relationships, and shows how they can be estimated from a simulation run. There are two types of estimators: an ordinary sample average, and a time-integrated (or time-weighted) sample average.

3.1 Time-Average Number in System *L*

Consider a queueing system over a period of time T, and let $L(t)$ denote the number of customers in the system at time t. A simulation of such a system is shown in Figure 6.

Let T_i denote the total time during $[0, T]$ in which the system contained exactly i customers. In Figure 6, it is seen that $T_0 = 3$, $T_1 = 12$, $T_2 = 4$, and $T_3 = 1$. (The line segments whose lengths total

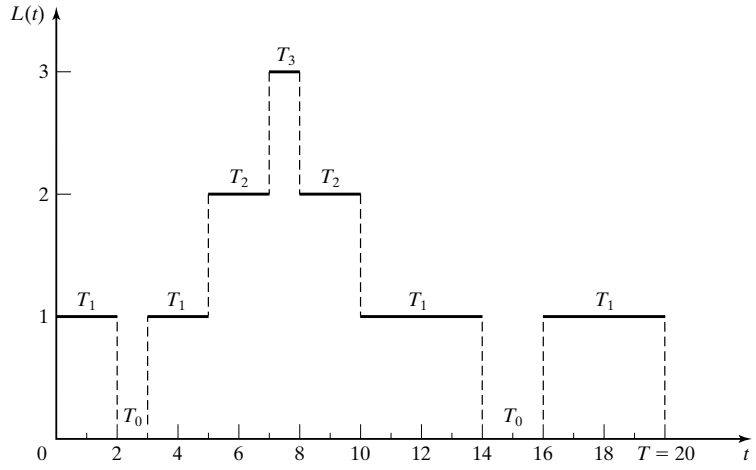

Figure 6 Number in system, $L(t)$, at time t.

$T_1 = 12$ are labeled "T_1" in Figure 6, etc.) In general, $\sum_{i=0}^{\infty} T_i = T$. The time-weighted-average number in a system is defined by

$$\widehat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \sum_{i=0}^{\infty} i\left(\frac{T_i}{T}\right) \tag{1}$$

For Figure 6, $\widehat{L} = [0(3) + 1(12) + 2(4) + 3(1)]/20 = 23/20 = 1.15$ customers. Notice that T_i/T is the proportion of time the system contains exactly i customers. The estimator $\widehat{L}$ is an example of a time-weighted average.

By considering Figure 6, it can be seen that the total area under the function $L(t)$ can be decomposed into rectangles of height i and length T_i. For example, the rectangle of area $3 \times T_3$ has base running from $t = 7$ to $t = 8$ (thus $T_3 = 1$); however, most of the rectangles are broken into parts, such as the rectangle of area $2 \times T_2$ which has part of its base between $t = 5$ and $t = 7$ and the remainder from $t = 8$ to $t = 10$ (thus $T_2 = 2 + 2 = 4$). It follows that the total area is given by $\sum_{i=0}^{\infty} iT_i = \int_0^T L(t)\,dt$, and, therefore, that

$$\widehat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \frac{1}{T}\int_0^T L(t)\,dt \tag{2}$$

The expressions in Equations (1) and (2) are always equal for any queueing system, regardless of the number of servers, the queue discipline, or any other special circumstances. Equation (2) justifies the terminology *time-integrated average*.

Many queueing systems exhibit a certain kind of long-run stability in terms of their average performance. For such systems, as time T gets large, the observed time-average number in the

system $\widehat{L}$ approaches a limiting value, say L, which is called the long-run time-average number in system—that is, with probability 1,

$$\widehat{L} = \frac{1}{T}\int_0^T L(t)\,dt \longrightarrow L \text{ as } T \longrightarrow \infty \tag{3}$$

The estimator $\widehat{L}$ is said to be strongly consistent for L. If simulation run length T is sufficiently long, the estimator $\widehat{L}$ becomes arbitrarily close to L. Unfortunately, for $T < \infty$, $\widehat{L}$ depends on the initial conditions at time 0.

Equations (2) and (3) can be applied to any subsystem of a queueing system as well as to the whole system. If $L_Q(t)$ denotes the number of customers waiting in queue, and T_i^Q denotes the total time during $[0, T]$ in which exactly i customers are waiting in queue, then

$$\widehat{L}_Q = \frac{1}{T}\sum_{i=0}^{\infty} iT_i^Q = \frac{1}{T}\int_0^T L_Q(t)\,dt \longrightarrow L_Q \text{ as } T \longrightarrow \infty \tag{4}$$

where $\widehat{L}_Q$ is the observed time-average number of customers waiting in queue from time 0 to time T and L_Q is the long-run time-average number waiting in queue.

Example 3

Suppose that Figure 6 represents a single-server queue—that is, a $G/G/1/N/K$ queueing system ($N \geq 3$, $K \geq 3$). Then the number of customers waiting in queue is given by $L_Q(t)$, defined by

$$L_Q(t) = \begin{cases} 0 & \text{if } L(t) = 0 \\ L(t) - 1 & \text{if } L(t) \geq 1 \end{cases}$$

and shown in Figure 7. Thus, $T_0^Q = 5 + 10 = 15$, $T_1^Q = 2 + 2 = 4$, and $T_2^Q = 1$. Therefore,

$$\widehat{L}_Q = \frac{0(15) + 1(4) + 2(1)}{20} = 0.3 \text{ customers}$$

3.2 Average Time Spent in System Per Customer *w*

If we simulate a queueing system for some period of time, say, T, then we can record the time each customer spends in the system during $[0, T]$, say $W_1, W_2, \ldots, W_N$, where N is the number of arrivals during $[0, T]$. The average time spent in system per customer, called the *average system time*, is given by the ordinary sample average

$$\widehat{w} = \frac{1}{N}\sum_{i=1}^{N} W_i \tag{5}$$

For stable systems, as $N \longrightarrow \infty$,

$$\widehat{w} \longrightarrow w \tag{6}$$

with probability 1, where w is called the *long-run average system time*.

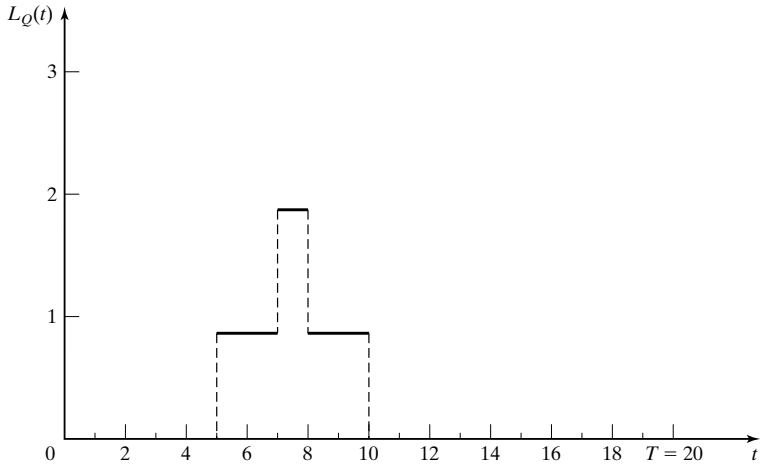

Figure 7 Number waiting in queue, $L_Q(t)$, at time t.

If the system under consideration is the queue alone, Equations (5) and (6) are written as

$$\widehat{w}_Q = \frac{1}{N}\sum_{i=1}^{N} W_i^Q \longrightarrow w_Q \quad \text{as } N \longrightarrow \infty \tag{7}$$

where W_i^Q is the total time customer i spends waiting in queue, $\widehat{w}_Q$ is the observed average time spent in queue (called *delay*), and w_Q is the long-run average delay per customer. The estimators $\widehat{w}$ and $\widehat{w}_Q$ are influenced by initial conditions at time 0 and the run length T, analogously to $\widehat{L}$.

Example 4

For the system history shown in Figure 6, $N = 5$ customers arrive, $W_1 = 2$, and $W_5 = 20 - 16 = 4$, but W_2, W_3, and W_4 cannot be computed unless more is known about the system. Assume that the system has a single server and a FIFO queue discipline. This implies that customers will depart from the system in the same order in which they arrived. Each jump upward of $L(t)$ in Figure 6 represents an arrival. Arrivals occur at times 0, 3, 5, 7, and 16. Similarly, departures occur at times 2, 8, 10, and 14. (A departure may or may not have occurred at time 20.) Under these assumptions, it is apparent that $W_2 = 8 - 3 = 5$, $W_3 = 10 - 5 = 5$, $W_4 = 14 - 7 = 7$, and therefore

$$\widehat{w} = \frac{2 + 5 + 5 + 7 + 4}{5} = \frac{23}{5} = 4.6 \text{ time units}$$

Thus, on the average, these customers spent 4.6 time units in the system. As for time spent in the waiting line, it can be computed that $W_1^Q = 0$, $W_2^Q = 0$, $W_3^Q = 8 - 5 = 3$, $W_4^Q = 10 - 7 = 3$, and

$W_5^Q = 0$; thus,

$$\widehat{w}_Q = \frac{0+0+3+3+0}{5} = 1.2 \text{ time units}$$

3.3 The Conservation Equation: $L = \lambda w$

For the system exhibited in Figure 6, there were $N = 5$ arrivals in $T = 20$ time units, and thus the observed arrival rate was $\widehat{\lambda} = N/T = 1/4$ customer per time unit. Recall that $\widehat{L} = 1.15$ and $\widehat{w} = 4.6$; hence, it follows that

$$\widehat{L} = \widehat{\lambda}\widehat{w} \tag{8}$$

This relationship between L, λ, and w is not coincidental; it holds for almost all queueing systems or subsystems regardless of the number of servers, the queue discipline, or any other special circumstances. Allowing $T \longrightarrow \infty$ and $N \longrightarrow \infty$, Equation (8) becomes

$$L = \lambda w \tag{9}$$

where $\widehat{\lambda} \longrightarrow \lambda$, and λ is the long-run average arrival rate. Equation (9) is called a conservation equation and is usually attributed to Little [1961]. It says that the average number of customers in the system at an arbitrary point in time is equal to the average number of arrivals per time unit, times the average time spent in the system. For Figure 6, there is one arrival every 4 time units (on average) and each arrival spends 4.6 time units in the system (on average), so at an arbitrary point in time there will be $(1/4)(4.6) = 1.15$ customers present (on average).

Equation (8) can also be derived by reconsidering Figure 6 in the following manner: Figure 8 shows system history, $L(t)$, exactly as in Figure 6, with each customer's time in the system, W_i, represented by a rectangle. This representation again assumes a single-server system with a FIFO queue discipline. The rectangles for the third and fourth customers are in two and three separate pieces, respectively. The ith rectangle has height 1 and length W_i for each $i = 1, 2, \ldots, N$. It follows that the total system time of all customers is given by the total area under the number-in-system function, $L(t)$; that is,

$$\sum_{i=1}^{N} W_i = \int_0^T L(t)\, dt \tag{10}$$

Therefore, by combining Equations (2) and (5) with $\widehat{\lambda} = N/T$, it follows that

$$\widehat{L} = \frac{1}{T}\int_0^T L(t)\, dt = \frac{N}{T}\frac{1}{N}\sum_{i=1}^{N} W_i = \widehat{\lambda}\widehat{w}$$

which is Little's equation (8). The intuitive and informal derivation presented here depended on the single-server FIFO assumptions, but these assumptions are not necessary. In fact, Equation (10),

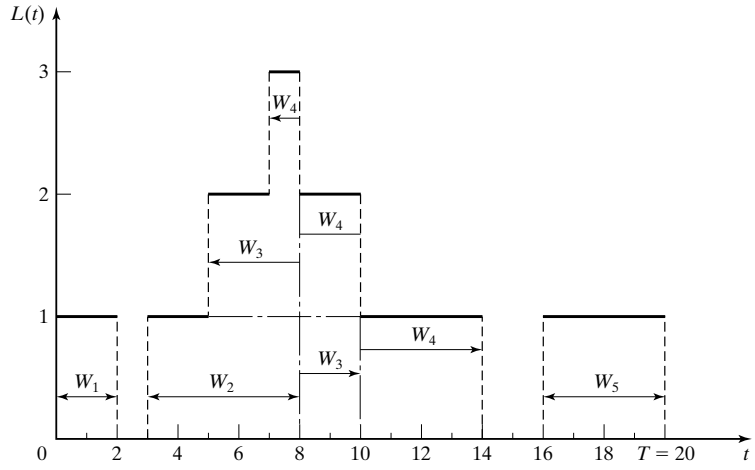

Figure 8 System times W_i for single-server FIFO system.

which was the key to the derivation, holds (at least approximately) in great generality, and thus so do Equations (8) and (9). Exercises 14 and 15 ask the reader to derive Equations (10) and (8) under different assumptions.

Technical note: If W_i is the system time for customer i during $[0, T]$, then Equation (10) and hence Equation (8) hold exactly. Some authors choose to define W_i as total system time for customer i; this change will affect the value of W_i only for those customers i who arrive before time T but do not depart until after time T (possibly customer 5 in Figure 8). With this change in definition, Equations (10) and (8) hold only approximately. Nevertheless, as $T \longrightarrow \infty$ and $N \longrightarrow \infty$, the error in Equation (8) decreases to zero, and, therefore, the conservation equation (9) for long-run measures of performance—namely, $L = \lambda w$—holds exactly.

3.4 Server Utilization

Server utilization is defined as the proportion of time that a server is busy. Observed server utilization, denoted by $\widehat{\rho}$, is defined over a specified time interval $[0, T]$. Long-run server utilization is denoted by ρ. For systems that exhibit long-run stability,

$$\widehat{\rho} \longrightarrow \rho \text{ as } T \longrightarrow \infty$$

Example 5

Per Figure 6 or 8, and assuming that the system has a single server, it can be seen that the server utilization is $\widehat{\rho} = (\text{total busy time})/T = (\sum_{i=1}^{\infty} T_i)/T = (T - T_0)/T = 17/20$.

Server utilization in $G/G/1/\infty/\infty$ queues

Consider any single-server queueing system with average arrival rate λ customers per time unit, average service time $E(S) = 1/\mu$ time units, and infinite queue capacity and calling population. Notice that $E(S) = 1/\mu$ implies that, when busy, the server is working at the rate μ customers per time unit, on the average; μ is called the *service rate*. The server alone is a subsystem that can be considered as a queueing system in itself; hence, the conservation Equation (9), $L = \lambda w$, can be applied to the server. For stable systems, the average arrival rate to the server, say λ_s, must be identical to the average arrival rate to the system, λ (certainly $\lambda_s \leq \lambda$—customers cannot be served faster than they arrive—but, if $\lambda_s < \lambda$, then the waiting line would tend to grow in length at an average rate of $\lambda - \lambda_s$ customers per time unit, and so we would have an unstable system). For the server subsystem, the average system time is $w = E(S) = \mu^{-1}$. The actual number of customers in the server subsystem is either 0 or 1, as shown in Figure 9 for the system represented by Figure 6. Hence, the average number in the server subsystem, $\widehat{L}_s$, is given by

$$\widehat{L}_s = \frac{1}{T}\int_0^T (L(t) - L_Q(t))\,dt = \frac{T - T_0}{T}$$

In this case, $\widehat{L}_s = 17/20 = \widehat{\rho}$. In general, for a single-server queue, the average number of customers being served at an arbitrary point in time is equal to server utilization. As $T \longrightarrow \infty$, $\widehat{L}_s = \widehat{\rho} \longrightarrow L_s = \rho$. Combining these results into $L = \lambda w$ for the server subsystem yields

$$\rho = \lambda E(S) = \frac{\lambda}{\mu} \tag{11}$$

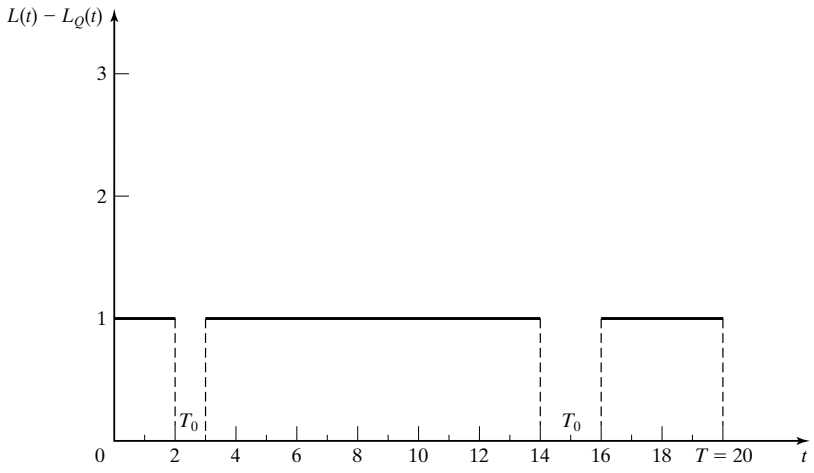

Figure 9 Number being served, $L(t) - L_Q(t)$, at time t.

that is, the long-run server utilization in a single-server queue is equal to the average arrival rate divided by the average service rate. For a single-server queue to be stable, the arrival rate λ must be less than the service rate μ:

$$\lambda < \mu$$

or

$$\rho = \frac{\lambda}{\mu} < 1 \tag{12}$$

If the arrival rate is greater than the service rate ($\lambda > \mu$), the server will eventually get further and further behind. After a time, the server will always be busy, and the waiting line will tend to grow in length at an average rate of ($\lambda - \mu$) customers per time unit, because departures will be occurring at rate μ per time unit. For stable single-server systems ($\lambda < \mu$ or $\rho < 1$), long-run measures of performance such as average queue length L_Q (and also L, w, and w_Q) are well defined and have meaning. For unstable systems ($\lambda > \mu$), long-run server utilization is 1, and long-run average queue length is infinite; that is,

$$\frac{1}{T}\int_0^T L_Q(t)\,dt \longrightarrow +\infty \text{ as } T \longrightarrow \infty$$

Similarly, $L = w = w_Q = \infty$. Therefore these long-run measures of performance are meaningless for unstable queues. The quantity λ/μ is also called the *offered load* and is a measure of the workload imposed on the system.

Server utilization in $G/G/c/\infty/\infty$ queues

Consider a queueing system with c identical servers in parallel. If an arriving customer finds more than one server idle, the customer chooses a server without favoring any particular server. (For example, the choice of server might be made at random.) Arrivals occur at rate λ from an infinite calling population, and each server works at rate μ customers per time unit. From Equation (9), $L = \lambda w$, applied to the server subsystem alone, an argument similar to the one given for a single server leads to the result that, for systems in statistical equilibrium, the average number of busy servers, say L_s, is given by

$$L_s = \lambda E(S) = \frac{\lambda}{\mu} \tag{13}$$

Clearly, $0 \leq L_s \leq c$. The long-run average server utilization is defined by

$$\rho = \frac{L_s}{c} = \frac{\lambda}{c\mu} \tag{14}$$

and so $0 \leq \rho \leq 1$. The utilization ρ can be interpreted as the proportion of time an arbitrary server is busy in the long run.

The maximum service rate of the $G/G/c/\infty/\infty$ system is $c\mu$, which occurs when all servers are busy. For the system to be stable, the average arrival rate λ must be less than the maximum service rate $c\mu$; that is, the system is stable if and only if

$$\lambda < c\mu \tag{15}$$

or, equivalently, if the offered load λ/μ is less than the number of servers c. If $\lambda > c\mu$, then arrivals are occurring, on the average, faster than the system can handle them, all servers will be continuously busy, and the waiting line will grow in length at an average rate of $(\lambda - c\mu)$ customers per time unit. Such a system is unstable, and the long-run performance measures (L, L_Q, w, and w_Q) are again meaningless for such systems.

Notice that Condition (15) generalizes Condition (12), and the equation for utilization for stable systems, Equation (14), generalizes Equation (11).

Equations (13) and (14) can also be applied when some servers work more than others; for example, when customers favor one server over others, or when certain servers serve customers only if all other servers are busy. In this case, the L_s given by Equation (13) is still the average number of busy servers, but ρ, as given by Equation (14), cannot be applied to an individual server. Instead, ρ must be interpreted as the average utilization of all servers.

Example 6

Customers arrive at random to a license bureau at a rate of $\lambda = 50$ customers per hour. Currently, there are 20 clerks, each serving $\mu = 5$ customers per hour on the average. Therefore the long-run, or steady-state, average utilization of a server, given by Equation (14), is

$$\rho = \frac{\lambda}{c\mu} = \frac{50}{20(5)} = 0.5$$

and the average number of busy servers is

$$L_s = \frac{\lambda}{\mu} = \frac{50}{5} = 10$$

Thus, in the long run, a typical clerk is busy serving customers only 50% of the time. The office manager asks whether the number of servers can be decreased. By Equation (15), it follows that, for the system to be stable, it is necessary for the number of servers to satisfy

$$c > \frac{\lambda}{\mu}$$

or $c > 50/5 = 10$. Thus, possibilities for the manager to consider include $c = 11$, or $c = 12$, or $c = 13, \ldots$. Notice that $c \geq 11$ guarantees long-run stability only in the sense that all servers, when busy, can handle the incoming work load (i.e., $c\mu > \lambda$) on average. The office manager could well desire to have more than the minimum number of servers ($c = 11$) because of other factors, such as customer delays and length of the waiting line. A stable queue can still have very long lines on average.

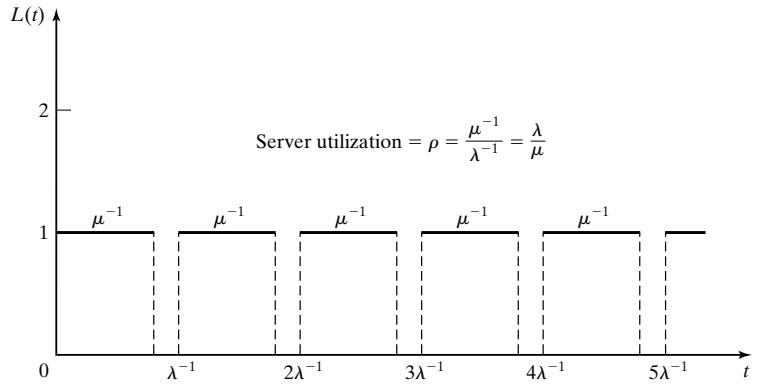

Figure 10 Deterministic queue ($D/D/1$).

Server utilization and system performance

As will be illustrated here and in later sections, system performance can vary widely for a given value of utilization, ρ. Consider a $G/G/1/\infty/\infty$ queue, that is, a single-server queue with arrival rate λ, service rate μ, and utilization $\rho = \lambda/\mu < 1$.

At one extreme, consider the $D/D/1$ queue, which has deterministic arrival and service times. Then all interarrival times $\{A_1, A_2, \dots\}$ are equal to $E(A) = 1/\lambda$, and all service times $\{S_1, S_2, \dots\}$ are equal to $E(S) = 1/\mu$. Assuming that a customer arrives to an empty system at time 0, the system evolves in a completely deterministic and predictable fashion, as shown in Figure 10. Observe that $L = \rho = \lambda/\mu, w = E(S) = \mu^{-1}$, and $L_Q = w_Q = 0$. By varying λ and μ, server utilization can assume any value between 0 and 1, yet there is never any line whatsoever. What, then, causes lines to build, if not a high server utilization? In general, it is the variability of interarrival and service times that causes lines to fluctuate in length.

Example 7

Consider a physician who schedules patients every 10 minutes and who spends S_i minutes with the ith patient, where

$$S_i = \begin{cases} 9 \text{ minutes with probability } 0.9 \\ 12 \text{ minutes with probability } 0.1 \end{cases}$$

Thus, arrivals are deterministic ($A_1 = A_2 = \cdots = \lambda^{-1} = 10$) but services are stochastic (or probabilistic), with mean and variance given by

$$E(S_i) = 9(0.9) + 12(0.1) = 9.3 \text{ minutes}$$

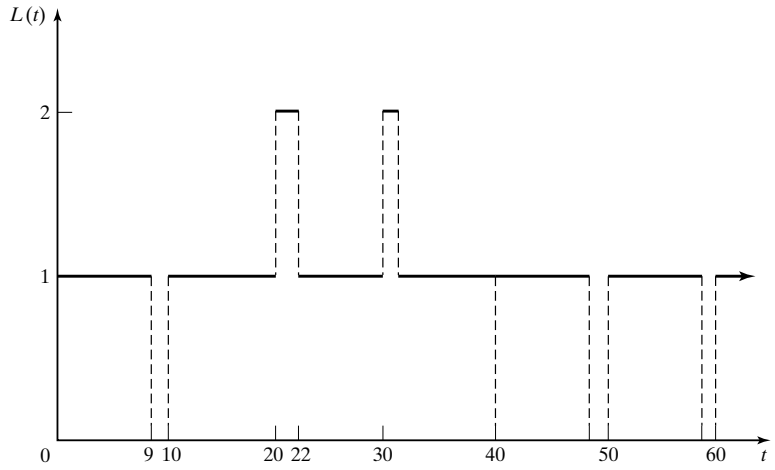

Figure 11 Number of patients in the doctor's office at time t.

and

$$\begin{aligned} V(S_i) &= E(S_i^2) - [E(S_i)]^2 \\ &= 9^2(0.9) + 12^2(0.1) - (9.3)^2 \\ &= 0.81 \text{ minutes}^2 \end{aligned}$$

Here, $\rho = \lambda/\mu = E(S)/E(A) = 9.3/10 = 0.93 < 1$, the system is stable, and the physician will be busy 93% of the time in the long run. In the short run, lines will not build up as long as patients require only 9 minutes of service, but, because of the variability in the service times, 10% of the patients will require 12 minutes, which in turn will cause a temporary line to form.

Suppose the system is simulated with service times, $S_1 = 9, S_2 = 12, S_3 = 9, S_4 = 9, S_5 = 9, \ldots$. Assuming that at time 0 a patient arrived to find the doctor idle and subsequent patients arrived precisely at times 10, 20, 30, ..., the system evolves as in Figure 11. The delays in queue are $W_1^Q = W_2^Q = 0, W_3^Q = 22 - 20 = 2, W_4^Q = 31 - 30 = 1, W_5^Q = 0$. The occurrence of a relatively long service time (here $S_2 = 12$) caused a waiting line to form temporarily. In general, because of the variability of the interarrival and service distributions, relatively small interarrival times and relatively large service times occasionally do occur, and these in turn cause lines to lengthen. Conversely, the occurrence of a large interarrival time or a small service time will tend to shorten an existing waiting line. The relationship between utilization, service and interarrival variability, and system performance will be explored in more detail in Section 4.

3.5 Costs in Queueing Problems

In many queueing situations, costs can be associated with various aspects of the waiting line or servers. Suppose that the system incurs a cost for each customer in the queue, say, at a rate of \$10 per hour per customer. If customer j spends W_j^Q hours in the queue, then $\sum_{j=1}^{N}(\$10 \cdot W_j^Q)$ is the total cost of the N customers who arrive during the simulation. Thus, the average cost per customer is

$$\sum_{j=1}^{N} \frac{\$10 \cdot W_j^Q}{N} = \$10 \cdot \widehat{w}_Q$$

by Equation (7). If $\widehat{\lambda}$ customers per hour arrive (on the average), the average cost per hour is

$$\left(\widehat{\lambda}\ \frac{\text{customers}}{\text{hour}}\right)\left(\frac{\$10 \cdot \widehat{w}_Q}{\text{customer}}\right) = \$10 \cdot \widehat{\lambda}\widehat{w}_Q = \$10 \cdot \widehat{L}_Q/\text{hour}$$

the last equality following by Equation (8). An alternative way to derive the average cost per hour is to consider Equation (2). If T_i^Q is the total time over the interval $[0, T]$ that the system contains exactly i customers, then \$10 iT_i^Q is the cost incurred by the system during the time exactly i customers are present. Thus, the total cost is $\sum_{i=1}^{\infty}$ ($\$10 \cdot iT_i^Q$), and the average cost per hour is

$$\sum_{i=1}^{\infty} \frac{\$10 \cdot iT_i^Q}{T} = \$10 \cdot \widehat{L}_Q/\text{ hour}$$

by Equation (2). In these cost expressions, $\widehat{L}_Q$ may be replaced by L_Q (if the long-run number in queue is known), or by L or $\widehat{L}$ (if costs are incurred while the customer is being served in addition to being delayed).

The server may also impose costs on the system. If a group of c parallel servers ($1 \leq c < \infty$) have utilization ρ, and each server imposes a cost of \$5 per hour while busy, the total server cost per hour is

$$\$5 \cdot c\rho$$

because $c\rho$ is the average number of busy servers. If server cost is imposed only when the servers are idle, then the server cost per hour would be

$$\$5 \cdot c(1 - \rho)$$

because $c(1 - \rho) = c - c\rho$ is the average number of idle servers. In many problems, two or more of these various costs are combined into a total cost. Such problems are illustrated by Exercises 1, 8, 12, and 19. In most cases, the objective is to minimize total costs (given certain constraints) by varying those parameters that are under management's control, such as the number of servers, the arrival rate, the service rate, and the system capacity.

4 Steady-State Behavior of Infinite-Population Markovian Models

This section presents steady-state results for a number of queueing models that can be solved mathematically. For the infinite-population models, the arrivals are assumed to follow a Poisson process with rate λ arrivals per time unit—that is, the interarrival times are assumed to be exponentially distributed with mean $1/\lambda$. Service times may be exponentially distributed (M) or arbitrarily (G). The queue discipline will be FIFO. Because of the exponential distribution assumptions on the arrival process, these models are called *Markovian models*.

A queueing system is said to be in *statistical equilibrium*, or *steady state*, if the probability that the system is in a given state is not time-dependent—that is,

$$P(L(t) = n) = P_n(t) = P_n$$

is independent of time t. Two properties—approaching statistical equilibrium from any starting state, and remaining in statistical equilibrium once it is reached—are characteristic of many stochastic models and, in particular, of all the systems studied in the following subsections. On the other hand, if an analyst were interested in the transient behavior of a queue over a relatively short period of time and were given some specific initial conditions (such as idle and empty), the results to be presented here would be inappropriate. A transient mathematical analysis or, more likely, a simulation model would be the chosen tool of analysis.

The mathematical models whose solutions are shown in the following subsections can be used to obtain approximate results even when the assumptions of the model do not strictly hold. These results may be considered as a rough guide to the behavior of the system. A simulation may then be used for a more refined analysis. However, it should be remembered that a mathematical analysis (when it is applicable) provides the true value of the model parameter (e.g., L), whereas a simulation analysis delivers a statistical estimate (e.g., $\widehat{L}$) of the parameter. On the other hand, for complex systems, a simulation model is often a more faithful representation than a mathematical model.

For the simple models studied here, the steady-state parameter L, the time-average number of customers in the system, can be computed as

$$L = \sum_{n=0}^{\infty} nP_n \tag{16}$$

where $\{P_n\}$ are the steady-state probabilities of finding n customers in the system (as defined in Table 2). As was discussed in Section 3 and was expressed in Equation (3), L can also be interpreted as a long-run measure of performance of the system. Once L is given, the other steady-state parameters can be computed readily from Little's equation (9) applied to the whole system and to the queue alone:

$$\begin{aligned} w &= \frac{L}{\lambda} \\ w_Q &= w - \frac{1}{\mu} \\ L_Q &= \lambda w_Q \end{aligned} \tag{17}$$

where λ is the arrival rate and μ is the service rate per server.

For the $M/G/c/\infty/\infty$ queues considered in this section to have a statistical equilibrium, a necessary and sufficient condition is that $\lambda/(c\mu) < 1$, where λ is the arrival rate, μ is the service rate of one server, and c is the number of parallel servers. For these unlimited capacity, infinite-calling-population models, it is assumed that the theoretical server utilization, $\rho = \lambda/(c\mu)$, satisfies $\rho < 1$. For models with finite system capacity or finite calling population, the quantity $\lambda/(c\mu)$ may assume any positive value.

4.1 Single-Server Queues with Poisson Arrivals and Unlimited Capacity: $M/G/1$

Suppose that service times have mean $1/\mu$ and variance σ^2 and that there is one server. If $\rho = \lambda/\mu < 1$, then the $M/G/1$ queue has a steady-state probability distribution with steady-state characteristics, as given in Table 3. In general, there is no simple expression for the steady-state probabilities $P_0, P_1, P_2, \ldots$. When $\lambda < \mu$, the quantity $\rho = \lambda/\mu$ is the server utilization, or long-run proportion of time the server is busy. As can be seen in Table 3, $1 - P_0 = \rho$ can also be interpreted as the steady-state probability that the system contains one or more customers. Notice also that $L - L_Q = \rho$ is the time-average number of customers being served.

Table 3 Steady-State Parameters of the $M/G/1$ Queue

ρ	$\dfrac{\lambda}{\mu}$
L	$\rho + \dfrac{\lambda^2(1/\mu^2 + \sigma^2)}{2(1-\rho)} = \rho + \dfrac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$
w	$\dfrac{1}{\mu} + \dfrac{\lambda(1/\mu^2 + \sigma^2)}{2(1-\rho)}$
w_Q	$\dfrac{\lambda(1/\mu^2 + \sigma^2)}{2(1-\rho)}$
L_Q	$\dfrac{\lambda^2(1/\mu^2 + \sigma^2)}{2(1-\rho)} = \dfrac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$
P_0	$1 - \rho$

Example 8

Customers arrive at a walk-in shoe repair shop apparently at random. It is assumed that arrivals occur according to a Poisson process at the rate $\lambda = 1.5$ per hour. Observation over several months has found that shoe repair times by the single worker take an average time of 30 minutes, with a standard deviation of 20 minutes. Thus the mean service time $1/\mu = 1/2$ hour, the service rate is $\mu = 2$ per hour and $\sigma^2 = (20)^2$ minutes^2 $= 1/9$ hour^2. The "customers" are the people needing shoe repair, and the appropriate model is the $M/G/1$ queue, because only the mean and variance of service times are

known, not their distribution. The proportion of time the worker is busy is $\rho = \lambda/\mu = 1.5/2 = 0.75$, and, by Table 3, the steady-state time average number of customers in the shop is

$$L = 0.75 + \frac{(1.5)^2[(0.5)^2 + 1/9]}{2(1 - 0.75)}$$
$$= 0.75 + 1.625 = 2.375 \text{ customers}$$

Thus, an observer who notes the state of the shoe repair system at arbitrary times would find an average of 2.375 customers (over the long run).

A closer look at the formulas in Table 3 reveals the source of the waiting lines and delays in an $M/G/1$ queue. For example, L_Q may be rewritten as

$$L_Q = \frac{\rho^2}{2(1 - \rho)} + \frac{\lambda^2\sigma^2}{2(1 - \rho)}$$

The first term involves only the ratio of the mean arrival rate λ to the mean service rate μ. As shown by the second term, if λ and μ are held constant, the average length of the waiting line (L_Q) depends on the variability, σ^2, of the service times. If two systems have identical mean service times and mean interarrival times, the one with the more variable service times (larger σ^2) will tend to have longer lines on the average. Intuitively, if service times are highly variable, then there is a high probability that a large service time will occur (say, much larger than the mean service time), and, when large service times do occur, there is a higher-than-usual tendency for lines to form and delays of customers to increase. (The reader should not confuse "steady state" with low variability or short lines; a system in steady-state or statistical equilibrium can be highly variable and can have long waiting lines.)

Example 9

There are two workers competing for a job. Able claims an average service time that is faster than Baker's, but Baker claims to be more consistent, even if not as fast. The arrivals occur according to a Poisson process at the rate $\lambda = 2$ per hour (1/30 per minute). Able's service statistics are an average service time of 24 minutes with a standard deviation of 20 minutes. Baker's service statistics are an average service time of 25 minutes, but a standard deviation of only 2 minutes. If the average length of the queue is the criterion for hiring, which worker should be hired? For Able, $\lambda = 1/30$ per minute, $1/\mu = 24$ minutes, $\sigma^2 = 20^2 = 400$ minutes^2, $\rho = \lambda/\mu = 24/30 = 4/5$, and the average queue length is computed as

$$L_Q = \frac{(1/30)^2[24^2 + 400]}{2(1 - 4/5)} = 2.711 \text{ customers}$$

For Baker, $\lambda = 1/30$ per minute, $1/\mu = 25$ minutes, $\sigma^2 = 2^2 = 4$ minutes^2, $\rho = 25/30 = 5/6$, and the average queue length is

$$L_Q = \frac{(1/30)^2[25^2 + 4]}{2(1 - 5/6)} = 2.097 \text{ customers}$$

Although working faster on the average, Able's greater service variability results in an average queue length about 30% greater than Baker's. On the basis of average queue length, L_Q, Baker wins. On the other hand, the proportion of arrivals who would find Able idle and thus experience no delay is $P_0 = 1 - \rho = 1/5 = 20\%$, but the proportion who would find Baker idle and thus experience no delay is $P_0 = 1 - \rho = 1/6 = 16.7\%$.

One case of the $M/G/1$ queue that is of special note occurs when service times are exponential, which we describe next.

The *M/M/1* queue. Suppose that service times in an $M/G/1$ queue are exponentially distributed, with mean $1/\mu$; then the variance as given by Equation $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$ is $\sigma^2 = 1/\mu^2$. The mean and standard deviation of the exponential distribution are equal, so the $M/M/1$ queue will often be a useful approximate model when service times have standard deviations approximately equal to their means. The steady-state parameters, given in Table 4, may be computed by substituting $\sigma^2 = 1/\mu^2$ into the formulas in Table 3. Alternatively, L may be computed by Equation (16) from the steady-state probabilities P_n given in Table 4, and then w, w_Q, and L_Q may be computed from the equation in (17). The student can show that the two expressions for each parameter are equivalent by substituting $\rho = \lambda/\mu$ into the right-hand side of each equation in Table 4.

Table 4 Steady-State Parameters of the $M/M/1$ Queue

L	$\dfrac{\lambda}{\mu - \lambda} = \dfrac{\rho}{1 - \rho}$
w	$\dfrac{1}{\mu - \lambda} = \dfrac{1}{\mu(1 - \rho)}$
w_Q	$\dfrac{\lambda}{\mu(\mu - \lambda)} = \dfrac{\rho}{\mu(1 - \rho)}$
L_Q	$\dfrac{\lambda^2}{\mu(\mu - \lambda)} = \dfrac{\rho^2}{1 - \rho}$
P_n	$\left(1 - \dfrac{\lambda}{\mu}\right)\left(\dfrac{\lambda}{\mu}\right)^n = (1 - \rho)\rho^n$

Example 10

Suppose that the interarrival times and service times at a single-chair unisex hair-styling shop have been shown to be exponentially distributed. The values of λ and μ are 2 per hour and 3 per hour, respectively—that is, the time between arrivals averages 1/2 hour, exponentially distributed, and the service time averages 20 minutes, also exponentially distributed. The server utilization and the probabilities for 0, 1, 2, 3, and 4 or more customers in the shop are computed as follows:

$$\rho = \frac{\lambda}{\mu} = \frac{2}{3}$$

$$P_0 = 1 - \frac{\lambda}{\mu} = \frac{1}{3}$$

$$P_1 = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right) = \frac{2}{9}$$

$$P_2 = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^2 = \frac{4}{27}$$

$$P_3 = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^3 = \frac{8}{81}$$

$$P_{\geq 4} = 1 - \sum_{n=0}^{3} P_n = 1 - \frac{1}{3} - \frac{2}{9} - \frac{4}{27} - \frac{8}{81} = \frac{16}{81}$$

From the calculations, the probability that the hair stylist is busy is $1 - P_0 = \rho = 0.67$; thus, the probability that the hair stylist is idle is 0.33. The time-average number of customers in the system is given by Table 4 as

$$L = \frac{\lambda}{\mu - \lambda} = \frac{2}{3 - 2} = 2 \text{ customers}$$

The average time an arrival spends in the system can be obtained from Table 4 or Equation (17) as

$$w = \frac{L}{\lambda} = \frac{2}{2} = 1 \text{ hour}$$

The average time the customer spends in the queue can be obtained from Equation (17) as

$$w_Q = w - \frac{1}{\mu} = 1 - \frac{1}{3} = \frac{2}{3} \text{ hour}$$

From Table 4, the time-average number in the queue is given by

$$L_Q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{4}{3(1)} = \frac{4}{3} \text{ customers}$$

Finally, notice that multiplying $w = w_Q + 1/\mu$ through by λ and using Little's equation (9) yields

$$L = L_Q + \frac{\lambda}{\mu} = \frac{4}{3} + \frac{2}{3} = 2 \text{ customers}$$

Example 11

For the $M/M/1$ queue with service rate $\mu = 10$ customers per hour, consider how L and w increase as the arrival rate λ increases from 5 to 8.64 by increments of 20%, and then to $\lambda = 10$.

λ	5.0	6.0	7.2	8.64	10.0
ρ	0.500	0.600	0.720	0.864	1.0
L	1.00	1.50	2.57	6.35	∞
w	0.20	0.25	0.36	0.73	∞

For any $M/G/1$ queue, if $\lambda/\mu \geq 1$, waiting lines tend to continually grow in length; the long-run measures of performance, L, w, w_Q, and L_Q are all infinite ($L = w = w_Q = L_Q = \infty$); and a steady-state probability distribution does not exist. As is shown here for $\lambda < \mu$, if ρ is close to 1, waiting lines and delays will tend to be long. Notice that the increase in average system time w and average number in system L is highly nonlinear as a function of ρ. For example, as λ increases by 20%, L increases first by 50% (from 1.00 to 1.50), then by 71% (to 2.57), and then by 147% (to 6.35).

Example 12

If arrivals are occurring at rate $\lambda = 10$ per hour, and management has a choice of two servers, one who works at rate $\mu_1 = 11$ customers per hour and the second at rate $\mu_2 = 12$ customers per hour, the respective utilizations are $\rho_1 = \lambda/\mu_1 = 10/11 = 0.909$ and $\rho_2 = \lambda/\mu_2 = 10/12 = 0.833$. If the $M/M/1$ queue is used as an approximate model, then, with the first server, the average number in the system would be, by Table 4,

$$L_1 = \frac{\rho_1}{1 - \rho_1} = 10$$

and, with the second server, the average number in the system would be

$$L_2 = \frac{\rho_2}{1 - \rho_2} = 5$$

Thus, an increase in service rate from 11 to 12 customers per hour, a mere 9.1% increase, would result in a decrease in average number in system from 10 to 5, which is a 50% decrease!

The effect of utilization and service variability

For any $M/G/1$ queue, if lines are too long, they can be reduced by decreasing the server utilization ρ or by decreasing the service time variability σ^2. These remarks hold for almost all queues, not just the $M/G/1$ queue. The utilization factor ρ can be reduced by decreasing the arrival rate λ, by increasing the service rate μ, or by increasing the number of servers, because, in general, $\rho = \lambda/(c\mu)$, where c is the number of parallel servers. The effect of additional servers will be studied in the following subsections.

The squared coefficient of variation (cv) of a positive random variable X is defined as

$$(\text{cv})^2 = \frac{V(X)}{[E(X)]^2}$$

It is a measure of the variability of a distribution. The larger its value, the more variable is the distribution relative to its expected value. For deterministic service times, $V(X) = 0$, so cv $= 0$. For Erlang service times of order k, $V(X) = 1/(k\mu^2)$ and $E(X) = 1/\mu$, so cv $= 1/\sqrt{k}$. For exponential service times at service rate μ, the mean service time is $E(X) = 1/\mu$ and the variance is $V(X) = 1/\mu^2$, so cv $= 1$. If service times have standard deviation greater than their mean (i.e., if cv > 1), then the hyperexponential distribution, which can achieve any desired coefficient of variation greater than 1, provides a good model. One occasion where it arises is given in Exercise 16.

The formula for L_Q for any $M/G/1$ queue can be rewritten in terms of the coefficient of variation by noticing that $(\text{cv})^2 = \sigma^2/(1/\mu)^2 = \sigma^2\mu^2$. Therefore,

$$\begin{aligned} L_Q &= \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)} \\ &= \frac{\rho^2(1+(\text{cv})^2)}{2(1-\rho)} \\ &= \left(\frac{\rho^2}{1-\rho}\right)\left(\frac{1+(\text{cv})^2}{2}\right) \end{aligned} \tag{18}$$

The first term, $\rho^2/(1-\rho)$, is L_Q for an $M/M/1$ queue. The second term, $\left(1+(\text{cv})^2\right)/2$, corrects the $M/M/1$ formula to account for a nonexponential service-time distribution. The formula for w_Q can be obtained from the corresponding $M/M/1$ formula by applying the same correction factor. These factors, $\rho^2/(1-\rho)$ and $\left(1+(\text{cv})^2\right)/2$, provide some insight into the relative impact of server utilization and server variability on queue congestion: L_Q explodes as $\rho \to 1$ while increasing linearly in $(\text{cv})^2$ for fixed ρ.

4.2 Multiserver Queue: $M/M/c/\infty/\infty$

Suppose that there are c channels operating in parallel. Each of these channels has an independent and identical exponential service-time distribution, with mean $1/\mu$. The arrival process is Poisson with rate λ. Arrivals will join a single queue and enter the first available service channel. The queueing system is shown in Figure 12. If the number in system is $n < c$, an arrival will enter an available channel. However, when $n \geq c$, a queue will build if arrivals occur.

The offered load is defined by λ/μ. If $\lambda \geq c\mu$, the arrival rate is greater than or equal to the maximum service rate of the system (the service rate when all servers are busy); thus, the system cannot handle the load put upon it, and therefore it has no statistical equilibrium. If $\lambda > c\mu$, the waiting line grows in length at the rate $\lambda - c\mu$ customers per time unit, on the average. Customers are entering the system at rate λ per time unit but are leaving the system at a maximum rate of $c\mu$ per time unit.

For the $M/M/c$ queue to have statistical equilibrium, the offered load must satisfy $\lambda/\mu < c$, in which case $\lambda/(c\mu) = \rho$, the server utilization. The steady-state parameters are listed in Table 5. Most

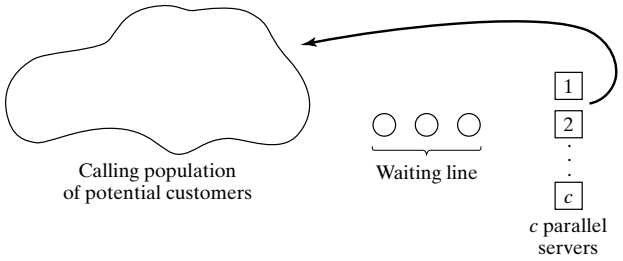

Figure 12 Multiserver queueing system.

Table 5 Steady-State Parameters for the $M/M/c$ Queue

ρ	$\dfrac{\lambda}{c\mu}$
P_0	$\left\{\left[\sum_{n=0}^{c-1}\dfrac{(\lambda/\mu)^n}{n!}\right]+\left[\left(\dfrac{\lambda}{\mu}\right)^c\left(\dfrac{1}{c!}\right)\left(\dfrac{c\mu}{c\mu-\lambda}\right)\right]\right\}^{-1}$
	$=\left\{\left[\sum_{n=0}^{c-1}\dfrac{(c\rho)^n}{n!}\right]+\left[(c\rho)^c\left(\dfrac{1}{c!}\right)\dfrac{1}{1-\rho}\right]\right\}^{-1}$
$P(L(\infty)\geq c)$	$\dfrac{(\lambda/\mu)^cP_0}{c!(1-\lambda/c\mu)}=\dfrac{(c\rho)^cP_0}{c!(1-\rho)}$
L	$c\rho+\dfrac{(c\rho)^{c+1}P_0}{c(c!)(1-\rho)^2}=c\rho+\dfrac{\rho P(L(\infty)\geq c)}{1-\rho}$
w	$\dfrac{L}{\lambda}$
w_Q	$w-\dfrac{1}{\mu}$
L_Q	$\lambda w_Q=\dfrac{(c\rho)^{c+1}P_0}{c(c!)(1-\rho)^2}=\dfrac{\rho P(L(\infty)\geq c)}{1-\rho}$
$L-L_Q$	$\dfrac{\lambda}{\mu}=c\rho$

of the measures of performance can be expressed fairly simply in terms of P_0, the probability that the system is empty, or $\sum_{n=c}^{\infty}P_n$, the probability that all servers are busy, denoted by $P(L(\infty)\geq c)$, where $L(\infty)$ is a random variable representing the number in system in statistical equilibrium (after a very long time). Thus, $P(L(\infty)=n)=P_n, n=0,1,2,\ldots$. The value of P_0 is necessary for computing

all the measures of performance, and the equation for P_0 is somewhat more complex than in the previous cases. However, the spreadsheet `QueueingTools.xls` available at `www.bcnn.net` performs these calculations for this and all queueing models described in the chapter.

The results in Table 5 simplify to those in Table 4 when $c = 1$, the case of a single server. Notice that the average number of busy servers, or the average number of customers being served, is given by the simple expression $L - L_Q = \lambda/\mu = c\rho$.

Example 13

Many early examples of queueing theory applied to practical problems concerning tool cribs. Attendants manage the tool cribs as mechanics, assumed to be from an infinite calling population, arrive for service. However, this example could as easily be customers arriving to the ticket counter at a movie theater. Assume Poisson arrivals at rate 2 mechanics per minute and exponentially distributed service times with mean 40 seconds.

Now, $\lambda = 2$ per minute, and $\mu = 60/40 = 3/2$ per minute. The offered load is greater than 1:

$$\frac{\lambda}{\mu} = \frac{2}{3/2} = \frac{4}{3} > 1$$

so more than one server is needed if the system is to have a statistical equilibrium. The requirement for steady state is that $c > \lambda/\mu = 4/3$. Thus at least $c = 2$ attendants are needed. The quantity 4/3 is the expected number of busy servers, and for $c \geq 2$, $\rho = 4/(3c)$ is the long-run proportion of time each server is busy. (What would happen if there were only $c = 1$ server?)

Let there be $c = 2$ attendants. First, P_0 is calculated as

$$P_0 = \left\{ \sum_{n=0}^{1} \frac{(4/3)^n}{n!} + \left(\frac{4}{3}\right)^2 \left(\frac{1}{2!}\right) \left[\frac{2(3/2)}{2(3/2) - 2}\right] \right\}^{-1}$$

$$= \left\{ 1 + \frac{4}{3} + \left(\frac{16}{9}\right)\left(\frac{1}{2}\right)(3) \right\}^{-1} = \left(\frac{15}{3}\right)^{-1} = \frac{1}{5} = 0.2$$

Next, the probability that all servers are busy is computed as

$$P(L(\infty) \geq 2) = \frac{(4/3)^2}{2!(1 - 2/3)} \left(\frac{1}{5}\right) = \left(\frac{8}{3}\right)\left(\frac{1}{5}\right) = \frac{8}{15} = 0.533$$

Thus, the time-average length of the waiting line of mechanics is

$$L_Q = \frac{(2/3)(8/15)}{1 - 2/3} = 1.07 \text{ mechanics}$$

and the time-average number in system is given by

$$L = L_Q + \frac{\lambda}{\mu} = \frac{16}{15} + \frac{4}{3} = \frac{12}{5} = 2.4 \text{ mechanics}$$

From Little's relationships, the average time a mechanic spends at the tool crib is

$$w = \frac{L}{\lambda} = \frac{2.4}{2} = 1.2 \text{ minutes}$$

and the average time spent waiting for an attendant is

$$w_Q = w - \frac{1}{\mu} = 1.2 - \frac{2}{3} = 0.533 \text{ minute}$$

An Approximation for the $M/G/c/\infty$ Queue

Recall that formulas for L_Q and w_Q for the $M/G/1$ queue can be obtained from the corresponding $M/M/1$ formulas by multiplying them by the correction factor $\left(1 + (\text{cv})^2\right)/2$, as in Equation (18). *Approximate* formulas for the $M/G/c$ queue can be obtained by applying the same correction factor to the $M/M/c$ formulas for L_Q and w_Q (no exact formula exists for $1 < c < \infty$). The nearer the cv is to 1, the better the approximation.

Example 14

Recall Example 13. Suppose that the service times for the mechanics at the tool crib are not exponentially distributed, but are known to have a standard deviation of 30 seconds. Then we have an $M/G/c$ model, rather than an $M/M/c$. The mean service time is 40 seconds, so the coefficient of variation of the service time is

$$\text{cv} = \frac{30}{40} = \frac{3}{4} < 1$$

Therefore, the accuracy of L_Q and w_Q can be improved by the correction factor

$$\frac{1 + (\text{cv})^2}{2} = \frac{1 + (3/4)^2}{2} = \frac{25}{32} = 0.78$$

For example, when there are $c = 2$ attendants,

$$L_Q = (0.78)(1.07) = 0.83 \text{ mechanics}$$

Notice that, because the coefficient of variation of the service time is less than 1, the congestion in the system, as measured by L_Q, is less than in the corresponding $M/M/2$ model.

The correction factor applies only to the formulas for L_Q and w_Q. Little's formula can then be used to calculate L and w. Unfortunately, there is no general method for correcting the steady-state probabilities, P_n.

When the Number of Servers is Infinite ($M/G/\infty/\infty$)

There are at least three situations in which it is appropriate to treat the number of servers as infinite:

1. when each customer is its own server—in other words, in a self-service system;
2. when service capacity far exceeds service demand, as in a so-called ample-server system; and
3. when we want to know how many servers are required so that customers will rarely be delayed.

The steady-state parameters for the $M/G/\infty$ queue are listed in Table 6. In the table, λ is the arrival rate of the Poisson arrival process, and $1/\mu$ is the expected service time of the general service-time distribution (including exponential, constant, or any other).

Table 6 Steady-State Parameters for the $M/G/\infty$ Queue

P_0	$e^{-\lambda/\mu}$
w	$\frac{1}{\mu}$
w_Q	0
L	$\frac{\lambda}{\mu}$
L_Q	0
P_n	$\frac{e^{-\lambda/\mu}(\lambda/\mu)^n}{n!}, n = 0, 1, \ldots$

Example 15

Prior to introducing their new, subscriber-only, online computer information service, The Connection must plan their system capacity in terms of the number of users that can be logged on simultaneously. If the service is successful, customers are expected to log on at a rate of $\lambda = 500$ per hour, according to a Poisson process, and stay connected for an average of $1/\mu = 180$ minutes (or 3 hours). In the real system, there will be an upper limit on simultaneous users, but, for planning purposes, The Connection can pretend that the number of simultaneous users is infinite. An $M/G/\infty$ model of the system implies that the expected number of simultaneous users is $L = \lambda/\mu = 500(3) = 1500$, so a capacity greater than 1500 is certainly required. To ensure providing adequate capacity 95% of the time, The Connection could allow the number of simultaneous users to be the smallest value c such that

$$P(L(\infty) \le c) = \sum_{n=0}^{c} P_n = \sum_{n=0}^{c} \frac{e^{-1500}(1500)^n}{n!} \ge 0.95$$

The capacity $c = 1564$ simultaneous users satisfies this requirement.

4.3 Multiserver Queues with Poisson Arrivals and Limited Capacity: $M/M/c/N/\infty$

Suppose that service times are exponentially distributed at rate μ, that there are c servers, and that the total system capacity is $N \geq c$ customers. If an arrival occurs when the system is full, that arrival is turned away and does not enter the system. As in the preceding section, suppose that arrivals occur randomly according to a Poisson process with rate λ arrivals per time unit. For any values of λ and μ such that $\rho \neq 1$, the $M/M/c/N$ queue has a statistical equilibrium with steady-state characteristics as given in Table 7 (formulas for the case $\rho = 1$ can be found in Hillier and Lieberman [2005]).

Table 7 Steady-State Parameters for the $M/M/c/N$ Queue

(N = System Capacity, $a = \lambda/\mu$, $\rho = \lambda/(c\mu)$)

P_0	$\left[1 + \sum_{n=1}^{c} \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=c+1}^{N} \rho^{n-c}\right]^{-1}$
P_N	$\frac{a^N}{c!c^{N-c}} P_0$
L_Q	$\frac{P_0 a^c \rho}{c!(1-\rho)^2}\left[1 - \rho^{N-c} - (N-c)\rho^{N-c}(1-\rho)\right]$
λ_e	$\lambda(1 - P_N)$
w_Q	$\frac{L_Q}{\lambda_e}$
w	$w_Q + \frac{1}{\mu}$
L	$\lambda_e w$

The effective arrival rate λ_e is defined as the mean number of arrivals per time unit who enter and remain in the system. For all systems, $\lambda_e \leq \lambda$; for the unlimited-capacity systems, $\lambda_e = \lambda$; but, for systems such as the present one, which turn customers away when full, $\lambda_e < \lambda$. The effective arrival rate is computed by

$$\lambda_e = \lambda(1 - P_N)$$

because $1 - P_N$ is the probability that a customer, upon arrival, will find space and be able to enter the system. When one is using Little's equations (17) to compute mean time spent in system w and in queue w_Q, λ must be replaced by λ_e.

Example 16

The unisex hair-styling shop described in Example 10 can hold only three customers: one in service, and two waiting. Additional customers are turned away when the system is full. The offered load is as previously determined, namely $\lambda/\mu = 2/3$.

To calculate the performance measures, first compute P_0:

$$P_0 = \left[1 + \frac{2}{3} + \frac{2}{3}\sum_{n=2}^{3}\left(\frac{2}{3}\right)^{n-1}\right]^{-1} = 0.415$$

The probability that there are three customers in the system (the system is full) is

$$P_N = P_3 = \frac{(2/3)^3}{1!1^2}P_0 = \frac{8}{65} = 0.123$$

Then, the average length of the queue (customers waiting for a haircut) is given by

$$L_Q = \frac{(27/65)(2/3)(2/3)}{(1-2/3)^2}\left[1 - (2/3)^2 - 2(2/3)^2(1-2/3)\right] = 0.431 \text{ customer}$$

Now, the effective arrival rate λ_e is given by

$$\lambda_e = 2\left(1 - \frac{8}{65}\right) = \frac{114}{65} = 1.754 \text{ customers per hour}$$

Therefore, from Little's equation, the expected time spent waiting in queue is

$$w_Q = \frac{L_Q}{\lambda_e} = \frac{28}{114} = 0.246 \text{ hour}$$

and the expected total time in the shop is

$$w = w_Q + \frac{1}{\mu} = \frac{66}{114} = 0.579 \text{ hour}$$

One last application of Little's equation gives the expected number of customers in the shop (in queue and getting a haircut) as

$$L = \lambda_e w = \frac{66}{65} = 1.015 \text{ customers}$$

Notice that $1 - P_0 = 0.585$ is the average number of customers being served or, equivalently, the probability that the single server is busy. Thus, the server utilization, or proportion of time the server is busy in the long run, is given by

$$1 - P_0 = \frac{\lambda_e}{\mu} = 0.585$$

The reader should compare these results to those of the unisex hair-styling shop before the capacity constraint was placed on the system. Specifically, in systems with limited capacity, the offered load λ/μ can assume any positive value and no longer equals the server utilization $\rho = \lambda_e/\mu$. Notice that server utilization decreases from 67% to 58.5% when the system imposes a capacity constraint.

5 Steady-State Behavior of Finite-Population Models ($M/M/c/K/K$)

In many practical problems, the assumption of an infinite calling population leads to invalid results because the calling population is, in fact, small. When the calling population is small, the presence of one or more customers in the system has a strong effect on the distribution of future arrivals, and the use of an infinite-population model can be misleading. Typical examples include a small group of machines that break down from time to time and require repair, or a small group of patients who are the responsibility of a staff of nurses. In the extreme case, if all the machines are broken, no new "arrivals" (breakdowns) of machines can occur; similarly, if all the patients have requested assistance, no arrival is possible. Contrast this to the infinite-population models, in which the arrival rate λ of customers to the system is assumed to be independent of the state of the system.

Consider a finite-calling-population model with K customers. The time between the end of one service visit and the next call for service for each member of the population is assumed to be exponentially distributed with mean $1/\lambda$ time units; service times are also exponentially distributed, with mean $1/\mu$ time units; there are c parallel servers, and system capacity is K, so that all arrivals remain for service. Such a system is depicted in Figure 13.

The steady-state parameters for this model are listed in Table 8. An electronic spreadsheet or a symbolic calculation program is useful for evaluating these complex formulas. For example, Figure 14 is a procedure written for the symbolic calculation program MATLAB to calculate the steady-state probabilities for the $M/M/c/K/K$ queue. The spreadsheet `QueueingTools.xls` available at `www.bcnn.net` also performs these calculations.

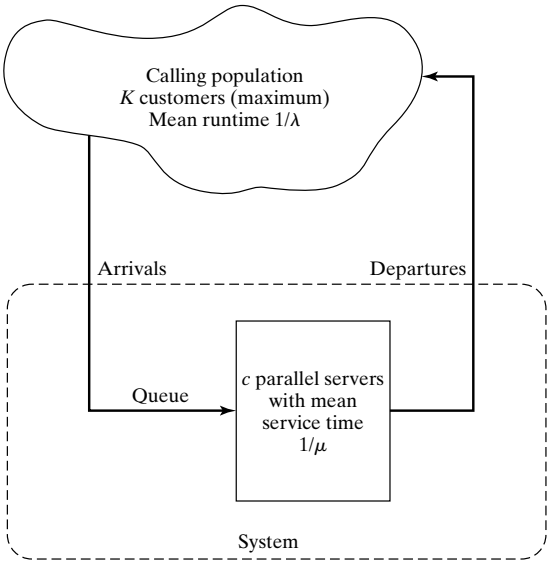

Figure 13 Finite-population queueing model.

Table 8 Steady-State Parameters for the $M/M/c/K/K$ Queue

P_0	$\left[\sum_{n=0}^{c-1}\binom{K}{n}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c}^{K}\frac{K!}{(K-n)!c!c^{n-c}}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}$
P_n	$\begin{cases}\binom{K}{n}\left(\frac{\lambda}{\mu}\right)^n P_0, & n = 0, 1, \ldots, c-1 \\ \frac{K!}{(K-n)!c!c^{n-c}}\left(\frac{\lambda}{\mu}\right)^n P_0, & n = c, c+1, \ldots, K\end{cases}$
L	$\sum_{n=0}^{K} nP_n$
L_Q	$\sum_{n=c+1}^{K}(n-c)P_n$
λ_e	$\sum_{n=0}^{K}(K-n)\lambda P_n$
w	L/λ_e
w_Q	L_Q/λ_e
ρ	$\frac{L-L_Q}{c} = \frac{\lambda_e}{c\mu}$

The effective arrival rate λ_e has several valid interpretations:

λ_e = long-run effective arrival rate of customers to the queue

= long-run effective arrival rate of customers entering service

= long-run rate at which customers exit from service

= long-run rate at which customers enter the calling population (and begin a new runtime)

= long-run rate at which customers exit from the calling population

Example 17

There are two workers who are responsible for 10 milling machines. The machines run on the average for 20 minutes, then require an average 5-minute service period, both times exponentially distributed. Therefore, $\lambda = 1/20$ and $\mu = 1/5$. Compute the various measures of performance for this system.

All of the performance measures depend on P_0, which is

$$\left[\sum_{n=0}^{2-1}\binom{10}{n}\left(\frac{5}{20}\right)^n + \sum_{n=c}^{10}\frac{10!}{(10-n)!2!2^{n-2}}\left(\frac{5}{20}\right)^n\right]^{-1} = 0.065$$

```
p = zeros(K+1,1);
% Note:
%   p(1) = lim_t->infty Pr{N(t)=0}
%   p(n+1) = lim_t->infty Pr{N(t)=n}, for n=1,2,...,K-1
%   p(K+1) = lim_t->infty Pr{N(t)=K}
crho = lambda/mu;
Kfac = factorial(K);
cfac = factorial(c);
% get p(1)
for n=0:c-1
    p(1)=p(1) + (Kfac/(factorial(n)*factorial(K-n)))*(crho^n);
end
for n=c:K
    p(1)=p(1) + (Kfac/((c^(n-c))*factorial(K-n)*cfac))*(crho^n);
end
p(1)=1/p(1);
% get p(n+1), 0 < n < c
for n=1:c-1
    p(n+1)=p(1) * (Kfac/(factorial(n)*factorial(K-n)))*(crho^n);
end
% get p(n+1), c <= n <= K
for n=c:K
    p(n+1)=p(1) * (Kfac/((c^(n-c))*factorial(K-n)*cfac))*(crho^n);
end
% return probability vector
mmcKK = p;
```

Figure 14 MATLAB procedure to calculate P_n for the $M/M/c/K/K$ queue.

From P_0, we can obtain the other P_n, from which we can compute the average number of machines waiting for service,

$$L_Q = \sum_{n=3}^{10} (n-2)P_n = 1.46 \text{ machines}$$

the effective arrival rate,

$$\lambda_e = \sum_{n=0}^{10} (10-n)\left(\frac{1}{20}\right) P_n = 0.342 \text{ machines/minute}$$

and the average waiting time in the queue,

$$w_Q = L_Q/\lambda_e = 4.27 \text{ minutes}$$

Similarly, we can compute the expected number of machines being serviced or waiting to be serviced,

$$L = \sum_{n=0}^{10} nP_n = 3.17 \text{ machines}$$

The average number of machines being serviced is given by

$$L - L_Q = 3.17 - 1.46 = 1.71 \text{ machines}$$

Each machine must be running, waiting to be serviced, or in service, so the average number of running machines is given by

$$K - L = 10 - 3.17 = 6.83 \text{ machines}$$

A question frequently asked is this: What will happen if the number of servers is increased or decreased? If the number of workers in this example increases to three ($c = 3$), then the time-average number of running machines increases to

$$K - L = 7.74 \text{ machines}$$

an increase of 0.91 machine, on the average.

Conversely, what happens if the number of servers decreases to one? Then the time-average number of running machines decreases to

$$K - L = 3.98 \text{ machines}$$

The decrease from two servers to one has resulted in a drop of nearly three machines running, on the average. Exercise 17 asks the reader to examine the effect on server utilization of adding or deleting one server.

Example 17 illustrates several general relationships that have been found to hold for almost all queues. If the number of servers is decreased, then delays, server utilization, and the probability of an arrival having to wait to begin service all increase.

6 Networks of Queues

In this chapter, we have emphasized the study of single queues of the $G/G/c/N/K$ type. However, many systems are naturally modeled as networks of single queues in which customers departing from one queue may be routed to another. Example 1 (see, in particular, Figure 3) and Example 2 (see Figure 5) are illustrations.

The study of mathematical models of networks of queues is beyond the scope of this chapter; see, for instance, Gross and Harris [1997], Nelson [1995], and Kleinrock [1976]. However, a few

fundamental principles are very useful for rough-cut modeling, perhaps prior to a simulation study. The following results assume a stable system with an infinite calling population and no limit on system capacity:

1. Provided that no customers are created or destroyed in the queue, then the departure rate out of a queue is the same as the arrival rate into the queue, over the long run.
2. If customers arrive to queue i at rate λ_i, and a fraction $0 \leq p_{ij} \leq 1$ of them are routed to queue j upon departure, then the arrival rate from queue i to queue j is $\lambda_i p_{ij}$, over the long run.
3. The overall arrival rate into queue j, λ_j, is the sum of the arrival rate from all sources. If customers arrive from outside the network at rate a_j, then

$$\lambda_j = a_j + \sum_{\text{all } i} \lambda_i p_{ij}$$

4. If queue j has $c_j < \infty$ parallel servers, each working at rate μ_j, then the long-run utilization of each server is

$$\rho_j = \frac{\lambda_j}{c_j \mu_j}$$

and $\rho_j < 1$ is required for the queue to be stable.
5. If, for each queue j, arrivals from outside the network form a Poisson process with rate a_j, and if there are c_j identical servers delivering exponentially distributed service times with mean $1/\mu_j$ (where c_j may be ∞), then, in steady state, queue j behaves like an $M/M/c_j$ queue with arrival rate $\lambda_j = a_j + \sum_{\text{all } i} \lambda_i p_{ij}$.

Example 18

Consider again the discount store described in Example 1 and shown in Figure 3. Suppose that customers arrive at the rate of 80 per hour and that, of those arrivals, 40% choose self-service; then, the arrival rate to service center 1 is $\lambda_1 = (80)(0.40) = 32$ per hour, and the arrival rate to service center 2 is $\lambda_2 = (80)(0.6) = 48$ per hour. Suppose that each of the $c_2 = 3$ clerks at service center 2 works at the rate $\mu_2 = 20$ customers per hour. Then the long-run utilization of the clerks is

$$\rho_2 = \frac{48}{(3)(20)} = 0.8$$

All customers must see the cashier at service center 3. The overall arrival rate to service center 3 is $\lambda_3 = \lambda_1 + \lambda_2 = 80$ per hour, regardless of the service rate at service center 1, because, over the long run, the departure rate out of each service center must be equal to the arrival rate into it. If the cashier works at rate $\mu_3 = 90$ per hour, then the utilization of the cashier is

$$\rho_3 = \frac{80}{90} = 0.89$$

7 Rough-cut Modeling: An Illustration

In this section we show how the tools described in this chapter can be used to do a rough-cut analysis prior to undertaking a more detailed simulation. Rough-cut modeling is useful in a number of ways: In some cases, the results of the rough-cut analysis are so compelling that there is no longer a need for the detailed simulation, saving a great deal of time and money. More typically, the rough-cut model provides the analyst with a better understanding of the system to be modeled—a kind of dress rehearsal prior to constructing a detailed and often complex simulation model. Further, the performance measures obtained at the rough-cut stage provide a sanity check on the simulation outputs, potentially averting erroneous conclusions due simply to mistakes in programming the simulation.

Example 19

At a driver's license branch office the manager receives many complaints about the long delays to renew licenses. To obtain quantitative support for more staff she brings in an operations analyst. A diagram of the facility is shown in Figure 15.

The branch is open from 8 A.M. to 4 P.M., with an historical average of 464 drivers per day being processed. All arrivals must first check in with one of two clerks. The manager has some time-study data collected on the clerks over several days and finds an average check-in time of 2 minutes with a standard deviation of about 0.4 minutes. After check in, 15% of the drivers need to take a written test that lasts approximately 20 minutes; good data are available on test time because the exams are time stamped when they are issued and returned. These multiple-choice tests are graded nearly instantly by an optical scanner. All arrivals must wait to have their picture taken and their license produced; this station can process about 60 drivers per hour, but the individual times are highly variable as some pictures have to be retaken.

The branch manager wants to know the relative impact on customer delay of adding a check-in clerk or adding a new photo station, and whether either of these changes will impact the number of chairs needed for drivers taking the written test (the 20 chairs she currently has have always been adequate).

This is an ideal application for simulation, but the analyst hired by the manager feels a quick queueing approximation could provide useful insight, as well as a check on the simulation model.

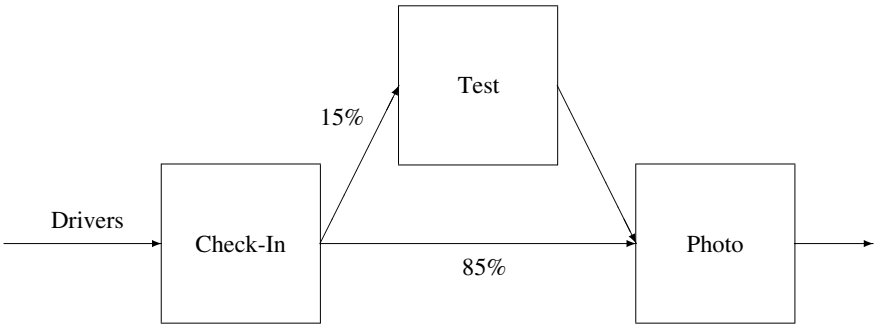

Figure 15 Customer flow in the driver's license branch office.

To do a rough-cut analysis, the analyst represents the license branch as a network of queues, with the check-in clerks being queue 1 (with $c_1 = 2$ servers, each working at rate $\mu_1 = 30$ drivers per hour $= 0.5$ drivers per minute), the written test station being queue 2 (with $c_2 = \infty$ servers, because any number of people can be taking the written test simultaneously, and mean service time $1/\mu_2 = 20$ minutes), and with the photo station being queue 3 (with $c_3 = 1$ server working at rate $\mu_3 = 1$ driver per minute). The analyst chooses to ignore the optical scanning of the tests because it happens so quickly. Although the branch manager says that arrivals are heavier around noon, for the purpose of a rough-cut analysis the analyst models the arrival rate as $464/8 = 58$ arrivals per hour throughout the day.

The arrival rates to each queue are as follows:

$$\lambda_1 = a_1 + \sum_{i=1}^{3} p_{i1}\lambda_i = 58 \text{ drivers per hour}$$

$$\lambda_2 = a_2 + \sum_{i=1}^{3} p_{i2}\lambda_i = (0.15)\lambda_1 \text{ drivers per hour}$$

$$\lambda_3 = a_3 + \sum_{i=1}^{3} p_{i3}\lambda_i = (1)\lambda_2 + (0.85)\lambda_1 \text{ drivers per hour}$$

Notice that arrivals from outside the network occur only at queue 1, so $a_1 = 50$ and $a_2 = a_3 = 0$. Solving this system of equations gives $\lambda_1 = \lambda_3 = 58$ and $\lambda_2 = 8.7$. The analyst chooses to work in minutes, and so converts these to $\lambda_1 = \lambda_3 = 0.97$ per minute and $\lambda_2 = 0.15$ per minute.

The analyst approximates the arrival process as Poisson—which is justified by the large population of potential customers who make (largely) independent decisions about when to arrive—and the service times at each queue as exponentially distributed. The service-time approximation is reasonable for queue 3, since the manager believes the picture-taking times are quite variable, and harmless for queue 2 because the results for infinite-server $M/G/\infty$ queues depend only on the service rate, not the distribution. However, the coefficient of variation of the check-in time is $\text{cv} = 0.4/2 = 0.2 < 1$, so the analyst decides to also apply the correction $(1 + \text{cv}^2)/2 = 0.52$ to the check-in results.

The analyst now has the necessary information to approximate each station in the license branch as an independent queue (another approximation, since in reality they are not independent). The check-in clerks are approximated as an $M/G/c_1$ queue, the testing station as an $M/G/\infty$ queue, and the photo station as an $M/M/c_3$ queue. Thus, under the current set-up, the check-in station is an $M/G/2$; using the formulas in Table 5 and multiplying by 0.52 gives $w_Q = 16.5$ minutes. If a third clerk is added, the waiting time in queue drops to 0.4 minute, a huge savings.

The current photo station can be modeled as an $M/M/1$ queue, giving $w_Q = 32.3$ minutes; adding a second photo station ($M/M/2$) causes the time in queue to drop to 0.3 minutes, a savings even larger than adding a clerk.

If desired, the testing station can be analyzed by using the results for an $M/G/\infty$ queue in Table 6. For instance, the expected number of people taking the test at any time is $L = \lambda_2/\mu_2 = 0.15/(1/20) = 3$, which will be unaffected by changes in the number of clerks or photo stations, since they do not change λ_2 or μ_2.

Should the manager expect to see precisely these results if either or both of the changes are implemented? No, because the model is rough in a number of ways: The queueing results are steady state, while the actual license facility is open only 8 hours per day, and does not have a constant arrival rate. The system was simplified by leaving out test scoring, employee breaks, etc. However, these results do provide a compelling case for the addition of one or more staff, and provide a baseline against which to check a detailed simulation model that includes all of the relevant features.

8 Summary

Queueing models have found widespread use in the analysis of service facilities, production and material-handling systems, telephone and communications systems, and many other situations where congestion or competition for scarce resources can occur. This chapter has introduced the basic concepts of queueing models and shown how simulation, and in some cases a mathematical analysis, can be used to estimate the performance measures of a system.

A simulation may be used to generate one or more artificial histories of a complex system. This simulation-generated data may, in turn, be used to estimate desired performance measures of the system. Commonly used performance measures, including L, L_Q, w, w_Q, ρ, and λ_e, were introduced, and formulas were given for their estimation from data.

When simulating any system that evolves over time, the analyst must decide whether transient behavior or steady-state performance is to be studied. Simple formulas exist for the steady-state behavior of some queues, but estimating steady-state performance measures from simulation-generated data requires recognizing and dealing with the possibly deleterious effect of the initial conditions on the estimators of steady-state performance. These estimators could be severely biased (either high or low) if the initial conditions are unrepresentative of steady state or if the simulation run length is too short.

Whether the analyst is interested in transient or in steady-state performance of a system, it should be recognized that the estimates obtained from a simulation of a stochastic queue are exactly that—estimates. Every such estimate contains random error, and a proper statistical analysis is required to assess the accuracy of the estimate.

In the last three sections of this chapter, it was shown that a number of simple models can be solved mathematically. Although the assumptions behind such models might not be met exactly in a practical application, these models can still be useful in providing a rough estimate of a performance measure. In many cases, models with exponentially distributed interarrival and service times will provide a conservative estimate of system behavior. For example, if the model predicts that average waiting time w will be 12.7 minutes, then average waiting time in the real system is likely to be less than 12.7 minutes. The conservative nature of exponential models arises because (a) performance measures, such as w and L, are generally increasing functions of the variance of interarrival times and service times (recall the $M/G/1$ queue), and (b) the exponential distribution is fairly highly variable, having its standard deviation always equal to its mean. Thus, if the arrival process or service mechanism of the real system is less variable than exponentially distributed interarrival or service times, it is likely that the average number in the system, L, and the average time spent in

system, w, will be less than what is predicted by the exponential model. Of course, if the interarrival and service times are *more* variable than exponential random variables, then the M/M queueing models could underestimate congestion.

An important application of mathematical queueing models is determining the minimum number of servers needed at a work station or service center. Quite often, if the arrival rate λ and the service rate μ are known or can be estimated, then the simple inequality $\lambda/(c\mu) < 1$ can be used to provide an initial estimate for the number of servers, c, at a work station. For a large system with many work stations, it could be quite time consuming to have to simulate every possibility $(c_1, c_2, \ldots)$ for the number of servers, c_i, at work station i. Thus, rough estimates from a bit of analysis could save a great deal of computer time and analysts' time.

Finally, the qualitative behavior of the simple exponential models of queueing carries over to more complex systems. In general, it is the variability of service times and the variability of the arrival process that causes waiting lines to build up and congestion to occur. For most systems, if the arrival rate increases, or if the service rate decreases, or if the variance of service times or interarrival times increases, then the system will become more congested. Congestion can be decreased by adding more servers or by reducing the mean and variability of service times. Simple queueing models can be a great aid in quantifying these relationships and in evaluating alternative system designs.

REFERENCES

COOPER, R. B. [1990], *Introduction to Queueing Theory*, 3d ed., George Washington University, Washington, DC.

DESCLOUX, A. [1962], *Delay Tables for Finite- and Infinite-Source Systems*, McGraw-Hill, New York.

GROSS, D., AND C. HARRIS [1997], *Fundamentals of Queueing Theory*, 3d ed., Wiley, New York.

HALL, R. W. [1991], *Queueing Methods: For Services and Manufacturing*, Prentice Hall, Englewood Cliffs, NJ.

HILLIER, F. S., AND G. J. LIEBERMAN [2005], *Introduction to Operations Research*, 8th ed., McGraw-Hill, New York.

KENDALL, D. G. [1953], "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Imbedded Markov Chains," *Annals of Mathematical Statistics*, Vol. 24, pp. 338–354.

KLEINROCK, L. [1976], *Queueing Systems, Vol. 2: Computer Applications*, Wiley, New York.

LITTLE, J. D. C. [1961], "A Proof for the Queueing Formula $L = \lambda w$," *Operations Research*, Vol. 16, pp. 651–665.

NELSON, B. L. [1995], *Stochastic Modeling*: *Analysis & Simulation*, Dover Publications, Mineola, NY.

WINSTON, W. L. [2004], *Operations Research: Applications and Algorithms*, 4th Edition, Duxbury Press, Pacific Grove, CA.

EXERCISES

1. A tool crib has exponential interarrival and service times and serves a very large group of mechanics. The mean time between arrivals is 4 minutes. It takes 3 minutes on the average for a tool-crib attendant to service a mechanic. The attendant is paid \$10 per hour and the mechanic is paid \$15 per hour. Would it be advisable to have a second tool-crib attendant?