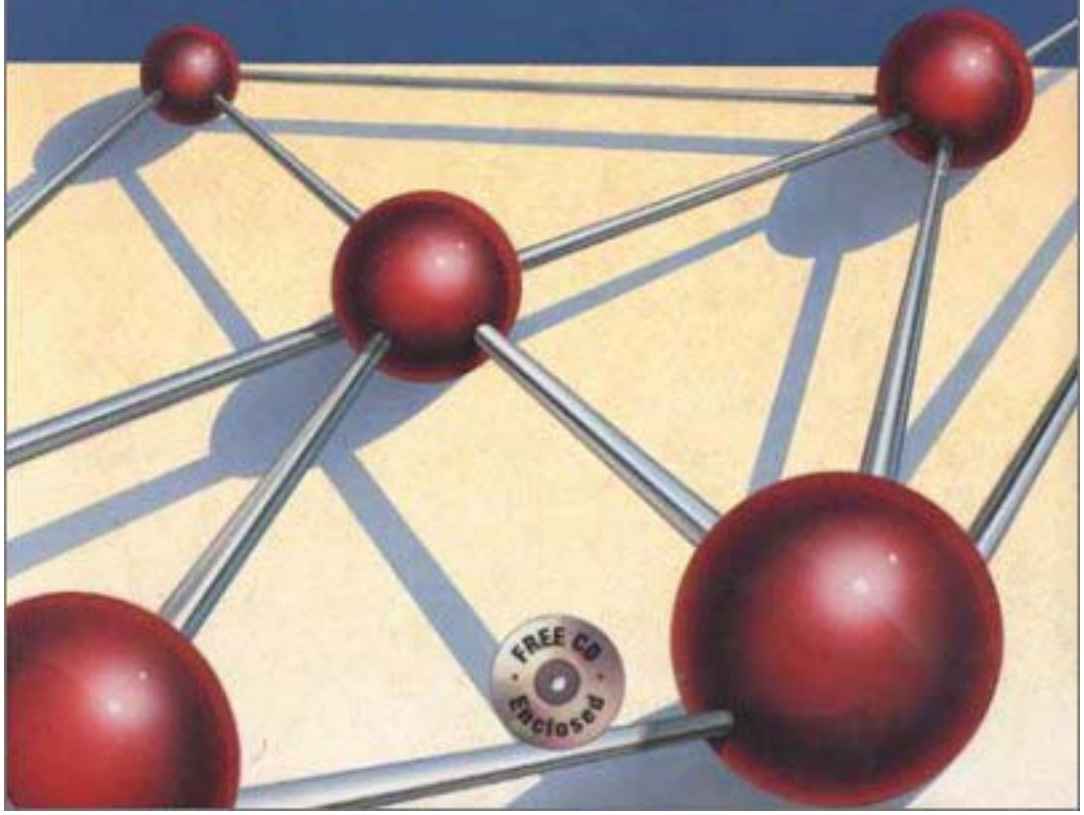
INTRODUCTION TO OPERATIONS RESEARCH

Seventh Edition

Hillier / Lieberman



17

Queueing Theory

Queues (waiting lines) are a part of everyday life. We all wait in queues to buy a movie ticket, make a bank deposit, pay for groceries, mail a package, obtain food in a cafeteria, start a ride in an amusement park, etc. We have become accustomed to considerable amounts of waiting, but still get annoyed by unusually long waits.

However, having to wait is not just a petty personal annoyance. The amount of time that a nation's populace wastes by waiting in queues is a major factor in both the quality of life there and the efficiency of the nation's economy. For example, before its dissolution, the U.S.S.R. was notorious for the tremendously long queues that its citizens frequently had to endure just to purchase basic necessities. Even in the United States today, it has been estimated that Americans spend 37,000,000,000 hours per year waiting in queues. If this time could be spent productively instead, it would amount to nearly 20 million person-years of useful work each year!

Even this staggering figure does not tell the whole story of the impact of causing excessive waiting. Great inefficiencies also occur because of other kinds of waiting than people standing in line. For example, making *machines* wait to be repaired may result in lost production. *Vehicles* (including ships and trucks) that need to wait to be unloaded may delay subsequent shipments. *Airplanes* waiting to take off or land may disrupt later travel schedules. Delays in *telecommunication* transmissions due to saturated lines may cause data glitches. Causing *manufacturing jobs* to wait to be performed may disrupt subsequent production. Delaying *service jobs* beyond their due dates may result in lost future business.

Queueing theory is the study of waiting in all these various guises. It uses *queueing models* to represent the various types of *queueing systems* (systems that involve queues of some kind) that arise in practice. Formulas for each model indicate how the corresponding queueing system should perform, including the average amount of waiting that will occur, under a variety of circumstances.

Therefore, these queueing models are very helpful for determining how to operate a queueing system in the most effective way. Providing too much service capacity to operate the system involves excessive costs. But not providing enough service capacity results in excessive waiting and all its unfortunate consequences. The models enable finding an appropriate balance between the cost of service and the amount of waiting.

After some general discussion, this chapter presents most of the more elementary queueing models and their basic results. Chapter 18 discusses how the information provided by queueing theory can be used to design queueing systems that minimize the total cost of service and waiting.

17.1 PROTOTYPE EXAMPLE

The emergency room of COUNTY HOSPITAL provides quick medical care for emergency cases brought to the hospital by ambulance or private automobile. At any hour there is always one doctor on duty in the emergency room. However, because of a growing tendency for emergency cases to use these facilities rather than go to a private physician, the hospital has been experiencing a continuing increase in the number of emergency room visits each year. As a result, it has become quite common for patients arriving during peak usage hours (the early evening) to have to wait until it is their turn to be treated by the doctor. Therefore, a proposal has been made that a second doctor should be assigned to the emergency room during these hours, so that two emergency cases can be treated simultaneously. The hospital's management engineer has been assigned to study this question.[1]

The management engineer began by gathering the relevant historical data and then projecting these data into the next year. Recognizing that the emergency room is a queueing system, she applied several alternative queueing theory models to predict the waiting characteristics of the system with one doctor and with two doctors, as you will see in the latter sections of this chapter (see Tables 17.2, 17.3, and 17.4).

17.2 BASIC STRUCTURE OF QUEUEING MODELS

The Basic Queueing Process

The basic process assumed by most queueing models is the following. *Customers* requiring service are generated over time by an *input source.* These customers enter the *queueing system* and join a *queue.* At certain times, a member of the queue is selected for service by some rule known as the *queue discipline.* The required service is then performed for the customer by the *service mechanism,* after which the customer leaves the queueing system. This process is depicted in Fig. 17.1.

Many alternative assumptions can be made about the various elements of the queueing process; they are discussed next.

Input Source (Calling Population)

One characteristic of the input source is its size. The *size* is the total number of customers that might require service from time to time, i.e., the total number of distinct potential customers. This population from which arrivals come is referred to as the **calling population.** The size may be assumed to be either *infinite* or *finite* (so that the input source also is said to be either *unlimited* or *limited*). Because the calculations are far easier for the infinite case, this assumption often is made even when the actual size is some rela-

[1]For one actual case study of this kind, see W. Blaker Bolling, "Queueing Model of a Hospital Emergency Room," *Industrial Engineering,* September 1972, pp. 26–31.

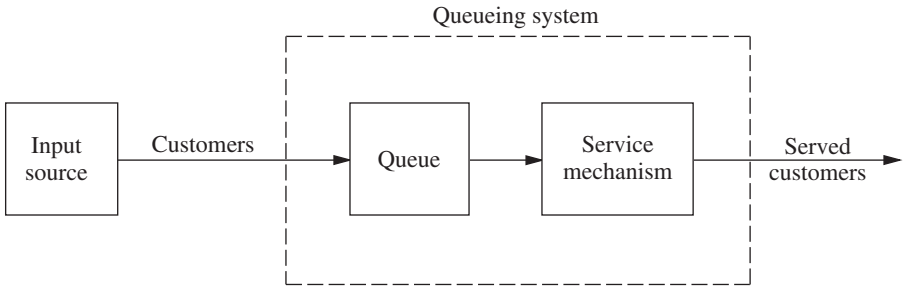


FIGURE 17.1
The basic queueing process.

tively large finite number; and it should be taken to be the implicit assumption for any queueing model that does not state otherwise. The finite case is more difficult analytically because the number of customers in the queueing system affects the number of potential customers outside the system at any time. However, the finite assumption must be made if the rate at which the input source generates new customers is significantly affected by the number of customers in the queueing system.

The statistical pattern by which customers are generated over time must also be specified. The common assumption is that they are generated according to a *Poisson process;* i.e., the number of customers generated until any specific time has a Poisson distribution. As we discuss in Sec. 17.4, this case is the one where arrivals to the queueing system occur randomly but at a certain fixed mean rate, regardless of how many customers already are there (so the *size* of the input source is *infinite*). An equivalent assumption is that the probability distribution of the time between consecutive arrivals is an *exponential* distribution. (The properties of this distribution are described in Sec. 17.4.) The time between consecutive arrivals is referred to as the **interarrival time.**

Any unusual assumptions about the behavior of arriving customers must also be specified. One example is *balking,* where the customer refuses to enter the system and is lost if the queue is too long.

Queue

The queue is where customers wait *before* being served. A queue is characterized by the maximum permissible number of customers that it can contain. Queues are called *infinite* or *finite,* according to whether this number is infinite or finite. The assumption of an *infinite queue* is the standard one for most queueing models, even for situations where there actually is a (relatively large) finite upper bound on the permissible number of customers, because dealing with such an upper bound would be a complicating factor in the analysis. However, for queueing systems where this upper bound is small enough that it actually would be reached with some frequency, it becomes necessary to assume a *finite queue.*

Queue Discipline

The queue discipline refers to the order in which members of the queue are selected for service. For example, it may be first-come-first-served, random, according to some priority procedure, or some other order. First-come-first-served usually is assumed by queueing models, unless it is stated otherwise.

Service Mechanism

The service mechanism consists of one or more *service facilities,* each of which contains one or more *parallel service channels,* called **servers.** If there is more than one service facility, the customer may receive service from a sequence of these (*service channels in series*). At a given facility, the customer enters one of the parallel service channels and is completely serviced by that server. A queueing model must specify the arrangement of the facilities and the number of servers (parallel channels) at each one. Most elementary models assume one service facility with either one server or a finite number of servers.

The time elapsed from the commencement of service to its completion for a customer at a service facility is referred to as the **service time** (or *holding time*). A model of a particular queueing system must specify the probability distribution of service times for each server (and possibly for different types of customers), although it is common to assume the *same* distribution for all servers (all models in this chapter make this assumption). The service-time distribution that is most frequently assumed in practice (largely because it is far more tractable than any other) is the *exponential* distribution discussed in Sec. 17.4, and most of our models will be of this type. Other important service-time distributions are the *degenerate* distribution (constant service time) and the *Erlang* (gamma) distribution, as illustrated by models in Sec. 17.7.

An Elementary Queueing Process

As we have already suggested, queueing theory has been applied to many different types of waiting-line situations. However, the most prevalent type of situation is the following: A single waiting line (which may be empty at times) forms in the front of a single service facility, within which are stationed one or more servers. Each customer generated by an input source is serviced by one of the servers, perhaps after some waiting in the queue (waiting line). The queueing system involved is depicted in Fig. 17.2.

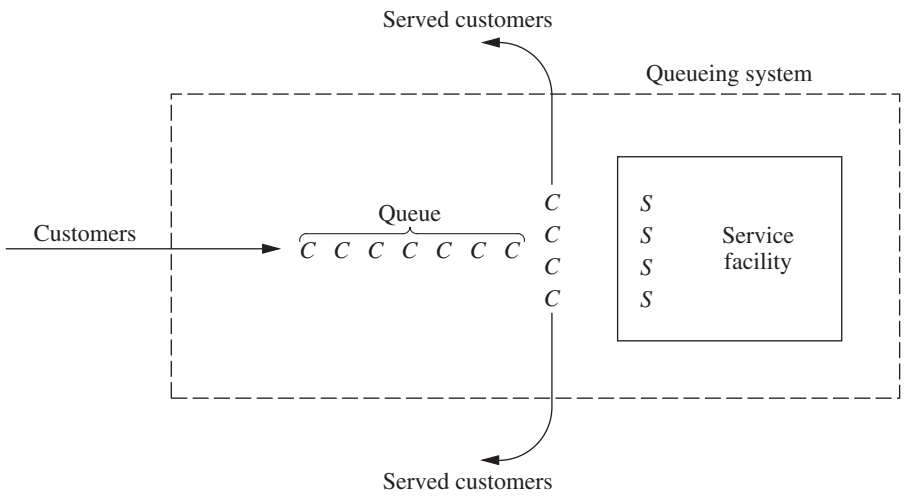


FIGURE 17.2
An elementary queueing system (each customer is indicated by a *C* and each server by an *S*).

Notice that the queueing process in the illustrative example of Sec. 17.1 is of this type. The input source generates customers in the form of emergency cases requiring medical care. The emergency room is the service facility, and the doctors are the servers.

A server need not be a single individual; it may be a group of persons, e.g., a repair crew that combines forces to perform simultaneously the required service for a customer. Furthermore, servers need not even be people. In many cases, a server can instead be a machine, a vehicle, an electronic device, etc. By the same token, the customers in the waiting line need not be people. For example, they may be items waiting for a certain operation by a given type of machine, or they may be cars waiting in front of a tollbooth.

It is not necessary that there actually be a physical waiting line forming in front of a physical structure that constitutes the service facility. The members of the queue may instead be scattered throughout an area, waiting for a server to come to them, e.g., machines waiting to be repaired. The server or group of servers assigned to a given area constitutes the service facility for that area. Queueing theory still gives the average number waiting, the average waiting time, and so on, because it is irrelevant whether the customers wait together in a group. The only essential requirement for queueing theory to be applicable is that changes in the number of customers waiting for a given service occur just as though the physical situation described in Fig. 17.2 (or a legitimate counterpart) prevailed.

Except for Sec. 17.9, all the queueing models discussed in this chapter are of the elementary type depicted in Fig. 17.2. Many of these models further assume that all *interarrival times* are independent and identically distributed and that all *service times* are independent and identically distributed. Such models conventionally are labeled as follows:

Distribution of service times

$$\_ / \_ / \_$$

Number of servers

Distribution of interarrival times,

where M = exponential distribution (Markovian), as described in Sec. 17.4,

D = degenerate distribution (constant times), as discussed in Sec. 17.7,

E_k = Erlang distribution (shape parameter = k), as described in Sec. 17.7,

G = general distribution (any arbitrary distribution allowed),[1] as discussed in Sec. 17.7.

For example, the $M/M/s$ model discussed in Sec. 17.6 assumes that both interarrival times and service times have an exponential distribution and that the number of servers is s (any positive integer). The $M/G/1$ model discussed again in Sec. 17.7 assumes that interarrival times have an exponential distribution, but it places no restriction on what the distribution of service times must be, whereas the number of servers is restricted to be exactly 1. Various other models that fit this labeling scheme also are introduced in Sec. 17.7.

[1]When we refer to interarrival times, it is conventional to replace the symbol G by GI = general independent distribution.

Terminology and Notation

Unless otherwise noted, the following standard terminology and notation will be used:

State of system = number of customers in queueing system.

Queue length = number of customers waiting for service to begin

= state of system *minus* number of customers being served.

$N(t)$ = number of customers in queueing system at time t ($t \geq 0$).

$P_n(t)$ = probability of exactly n customers in queueing system at time t, given number at time 0.

s = number of servers (parallel service channels) in queueing system.

λ_n = mean arrival rate (expected number of arrivals per unit time) of new customers when n customers are in system.

μ_n = mean service rate for overall system (expected number of customers completing service per unit time) when n customers are in system. *Note:* μ_n represents *combined* rate at which all *busy* servers (those serving customers) achieve service completions.

λ, μ, ρ = see following paragraph.

When λ_n is a constant for all n, this constant is denoted by λ. When the mean service rate *per busy server* is a constant for all $n \geq 1$, this constant is denoted by μ. (In this case, $\mu_n = s\mu$ when $n \geq s$, that is, when all s servers are busy.) Under these circumstances, $1/\lambda$ and $1/\mu$ are the *expected interarrival time* and the *expected service time,* respectively. Also, $\rho = \lambda/(s\mu)$ is the **utilization factor** for the service facility, i.e., the expected fraction of time the individual servers are busy, because $\lambda/(s\mu)$ represents the fraction of the system's service capacity ($s\mu$) that is being *utilized* on the average by arriving customers (λ).

Certain notation also is required to describe *steady-state* results. When a queueing system has recently begun operation, the state of the system (number of customers in the system) will be greatly affected by the initial state and by the time that has since elapsed. The system is said to be in a **transient condition.** However, after sufficient time has elapsed, the state of the system becomes essentially independent of the initial state and the elapsed time (except under unusual circumstances).[1] The system has now essentially reached a **steady-state condition,** where the probability distribution of the state of the system remains the same (the *steady-state* or *stationary* distribution) over time. Queueing theory has tended to focus largely on the steady-state condition, partially because the transient case is more difficult analytically. (Some transient results exist, but they are generally beyond the technical scope of this book.) The following notation assumes that the system is in a *steady-state condition:*

P_n = probability of exactly n customers in queueing system.

L = expected number of customers in queueing system $= \sum_{n=0}^{\infty} nP_n$.

[1]When λ and μ are defined, these unusual circumstances are that $\rho \geq 1$, in which case the state of the system tends to grow continually larger as time goes on.

L_q = expected queue length (excludes customers being served) $= \sum_{n=s}^{\infty} (n - s)P_n$.

$\mathcal{W}$ = waiting time in system (includes service time) for each individual customer.

$W = E(\mathcal{W})$.

$\mathcal{W}_q$ = waiting time in queue (excludes service time) for each individual customer.

$W_q = E(\mathcal{W}_q)$.

Relationships between L, W, L_q, and W_q

Assume that λ_n is a constant λ for all n. It has been proved that in a steady-state queueing process,

$$L = \lambda W.$$

(Because John D. C. Little[1] provided the first rigorous proof, this equation sometimes is referred to as **Little's formula.**) Furthermore, the same proof also shows that

$$L_q = \lambda W_q.$$

If the λ_n are not equal, then λ can be replaced in these equations by $\bar{\lambda}$, the *average* arrival rate over the long run. (We shall show later how $\bar{\lambda}$ can be determined for some basic cases.)

Now assume that the mean service time is a constant, $1/\mu$ for all $n \geq 1$. It then follows that

$$W = W_q + \frac{1}{\mu}.$$

These relationships are extremely important because they enable all four of the fundamental quantities—L, W, L_q, and W_q—to be immediately determined as soon as one is found analytically. This situation is fortunate because some of these quantities often are much easier to find than others when a queueing model is solved from basic principles.

17.3 EXAMPLES OF REAL QUEUEING SYSTEMS

Our description of queueing systems in the preceding section may appear relatively abstract and applicable to only rather special practical situations. On the contrary, queueing systems are surprisingly prevalent in a wide variety of contexts. To broaden your horizons on the applicability of queueing theory, we shall briefly mention various examples of real queueing systems.

One important class of queueing systems that we all encounter in our daily lives is **commercial service systems,** where outside customers receive service from commercial organizations. Many of these involve person-to-person service at a fixed location, such as a barber shop (the barbers are the servers), bank teller service, checkout stands at a grocery store, and a cafeteria line (service channels in series). However, many others do not,

[1] J. D. C. Little, "A Proof for the Queueing Formula: $L = \lambda W$," *Operations Research,* **9**(3): 383–387, 1961; also see S. Stidham, Jr., "A Last Word on $L = \lambda W$," *Operations Research,* **22**(2): 417–421, 1974.

such as home appliance repairs (the server travels to the customers), a vending machine (the server is a machine), and a gas station (the cars are the customers).

Another important class is **transportation service systems.** For some of these systems the vehicles are the customers, such as cars waiting at a tollbooth or traffic light (the server), a truck or ship waiting to be loaded or unloaded by a crew (the server), and airplanes waiting to land or take off from a runway (the server). (An unusual example of this kind is a parking lot, where the cars are the customers and the parking spaces are the servers, but there is no queue because arriving customers go elsewhere to park if the lot is full.) In other cases, the vehicles, such as taxicabs, fire trucks, and elevators, are the servers.

In recent years, queueing theory probably has been applied most to **internal service systems,** where the customers receiving service are *internal* to the organization. Examples include materials-handling systems, where materials-handling units (the servers) move loads (the customers); maintenance systems, where maintenance crews (the servers) repair machines (the customers); and inspection stations, where quality control inspectors (the servers) inspect items (the customers). Employee facilities and departments servicing employees also fit into this category. In addition, machines can be viewed as servers whose customers are the jobs being processed. A related example is a computer laboratory, where each computer is viewed as the server.

There is now growing recognition that queueing theory also is applicable to **social service systems.** For example, a judicial system is a queueing network, where the courts are service facilities, the judges (or panels of judges) are the servers, and the cases waiting to be tried are the customers. A legislative system is a similar queueing network, where the customers are the bills waiting to be processed. Various health-care systems also are queueing systems. You already have seen one example in Sec. 17.1 (a hospital emergency room), but you can also view ambulances, x-ray machines, and hospital beds as servers in their own queueing systems. Similarly, families waiting for low- and moderate-income housing, or other social services, can be viewed as customers in a queueing system.

Although these are four broad classes of queueing systems, they still do not exhaust the list. In fact, queueing theory first began early in this century with applications to telephone engineering (the founder of queueing theory, A. K. Erlang, was an employee of the Danish Telephone Company in Copenhagen), and telephone engineering still is an important application. Furthermore, we all have our own personal queues—homework assignments, books to be read, and so forth. However, these examples are sufficient to suggest that queueing systems do indeed pervade many areas of society.

17.4 THE ROLE OF THE EXPONENTIAL DISTRIBUTION

The operating characteristics of queueing systems are determined largely by two statistical properties, namely, the probability distribution of *interarrival times* (see "Input Source" in Sec. 17.2) and the probability distribution of *service times* (see "Service Mechanism" in Sec. 17.2). For real queueing systems, these distributions can take on almost any form. (The only restriction is that negative values cannot occur.) However, to formulate a queueing theory *model* as a representation of the real system, it is necessary to specify the assumed form of each of these distributions. To be useful, the assumed form should be *sufficiently realistic* that the model provides *reasonable predictions* while, at the same time,

being *sufficiently simple* that the model is *mathematically tractable.* Based on these considerations, the most important probability distribution in queueing theory is the *exponential distribution.*

Suppose that a random variable T represents either interarrival or service times. (We shall refer to the occurrences marking the end of these times—arrivals or service completions—as *events.*) This random variable is said to have an *exponential distribution with parameter* α if its probability density function is

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0, \end{cases}$$

as shown in Fig. 17.3. In this case, the cumulative probabilities are

$$\begin{aligned} P\{T \leq t\} &= 1 - e^{-\alpha t} \\ P\{T > t\} &= e^{-\alpha t} \end{aligned} \qquad (t \geq 0),$$

and the expected value and variance of T are, respectively,

$$E(T) = \frac{1}{\alpha},$$

$$\text{var}(T) = \frac{1}{\alpha^2}.$$

What are the implications of assuming that T has an exponential distribution for a queueing model? To explore this question, let us examine six key properties of the exponential distribution.

Property 1: $f_T(t)$ is a strictly *decreasing* function of t ($t \geq 0$).

One consequence of Property 1 is that

$$P\{0 \leq T \leq \Delta t\} > P\{t \leq T \leq t + \Delta t\}$$

for any strictly positive values of Δt and t. [This consequence follows from the fact that these probabilities are the area under the $f_T(t)$ curve over the indicated interval of length Δt, and the average height of the curve is less for the second probability than for the first.]

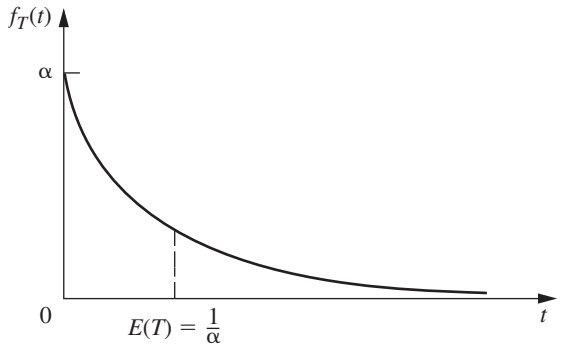


FIGURE 17.3
Probability density function for the exponential distribution.

Therefore, it is not only possible but also relatively likely that T will take on a small value near zero. In fact,

$$P\left\{0 \leq T \leq \frac{1}{2}\frac{1}{\alpha}\right\} = 0.393$$

whereas

$$P\left\{\frac{1}{2}\frac{1}{\alpha} \leq T \leq \frac{3}{2}\frac{1}{\alpha}\right\} = 0.383,$$

so that the value T takes on is more likely to be "small" [i.e., less than half of $E(T)$] than "near" its expected value [i.e., no further away than half of $E(T)$], even though the second interval is twice as wide as the first.

Is this really a reasonable property for T in a queueing model? If T represents *service times,* the answer depends upon the general nature of the service involved, as discussed next.

If the service required is essentially identical for each customer, with the server always performing the same sequence of service operations, then the actual service times tend to be near the expected service time. Small deviations from the mean may occur, but usually because of only minor variations in the efficiency of the server. A small service time far below the mean is essentially impossible, because a certain minimum time is needed to perform the required service operations even when the server is working at top speed. The exponential distribution clearly does not provide a close approximation to the service-time distribution for this type of situation.

On the other hand, consider the type of situation where the specific tasks required of the server differ among customers. The broad nature of the service may be the same, but the specific type and amount of service differ. For example, this is the case in the County Hospital emergency room problem discussed in Sec. 17.1. The doctors encounter a wide variety of medical problems. In most cases, they can provide the required treatment rather quickly, but an occasional patient requires extensive care. Similarly, bank tellers and grocery store checkout clerks are other servers of this general type, where the required service is often brief but must occasionally be extensive. An exponential service-time distribution would seem quite plausible for this type of service situation.

If T represents *interarrival times,* Property 1 rules out situations where potential customers approaching the queueing system tend to postpone their entry if they see another customer entering ahead of them. On the other hand, it is entirely consistent with the common phenomenon of arrivals occurring "randomly," described by subsequent properties. Thus, when arrival times are plotted on a time line, they sometimes have the appearance of being clustered with occasional large gaps separating clusters, because of the substantial probability of small interarrival times and the small probability of large interarrival times, but such an irregular pattern is all part of true randomness.

Property 2: Lack of memory.

This property can be stated mathematically as

$$P\{T > t + \Delta t \mid T > \Delta t\} = P\{T > t\}$$

for any positive quantities t and Δt. In other words, the probability distribution of the *remaining* time until the event (arrival or service completion) occurs always is the same, regardless of how much time (Δt) already has passed. In effect, the process "forgets" its history. This surprising phenomenon occurs with the exponential distribution because

$$\begin{aligned} P\{T > t + \Delta t \mid T > \Delta t\} &= \frac{P\{T > \Delta t, T > t + \Delta t\}}{P\{T > \Delta t\}} \\ &= \frac{P\{T > t + \Delta t\}}{P\{T > \Delta t\}} \\ &= \frac{e^{-\alpha(t+\Delta t)}}{e^{-\alpha \Delta t}} \\ &= e^{-\alpha t} \\ &= P\{T > t\}. \end{aligned}$$

For *interarrival times,* this property describes the common situation where the time until the next arrival is completely uninfluenced by when the last arrival occurred. For *service times,* the property is more difficult to interpret. We should not expect it to hold in a situation where the server must perform the same fixed sequence of operations for each customer, because then a long elapsed service should imply that probably little remains to be done. However, in the type of situation where the required service operations differ among customers, the mathematical statement of the property may be quite realistic. For this case, if considerable service has already elapsed for a customer, the only implication may be that this particular customer requires more extensive service than most.

Property 3: The *minimum* of several independent exponential random variables has an exponential distribution.

To state this property mathematically, let $T_1, T_2, \ldots, T_n$ be *independent* exponential random variables with parameters $\alpha_1, \alpha_2, \ldots, \alpha_n$, respectively. Also let U be the random variable that takes on the value equal to the *minimum* of the values actually taken on by $T_1, T_2, \ldots, T_n$; that is,

$$U = \min\{T_1, T_2, \ldots, T_n\}.$$

Thus, if T_i represents the time until a particular kind of event occurs, then U represents the time until the *first* of the n different events occurs. Now note that for any $t \geq 0$,

$$\begin{aligned} P\{U > t\} &= P\{T_1 > t, T_2 > t, \ldots, T_n > t\} \\ &= P\{T_1 > t\}P\{T_2 > t\} \cdots P\{T_n > t\} \\ &= e^{-\alpha_1 t} e^{-\alpha_2 t} \cdots e^{\alpha_n t} \\ &= \exp\left(-\sum_{i=1}^{n} \alpha_i t\right), \end{aligned}$$

so that U indeed has an exponential distribution with parameter

$$\alpha = \sum_{i=1}^{n} \alpha_i.$$

This property has some implications for interarrival times in queueing models. In particular, suppose that there are several (n) *different* types of customers, but the interarrival

times for *each* type (type i) have an exponential distribution with parameter α_i ($i = 1, 2, \ldots, n$). By Property 2, the *remaining* time from any specified instant until the next arrival of a customer of type i has this same distribution. Therefore, let T_i be this remaining time, measured from the instant a customer of *any* type arrives. Property 3 then tells us that U, the interarrival times for the queueing system as a whole, has an exponential distribution with parameter α defined by the last equation. As a result, you can choose to ignore the distinction between customers and still have exponential interarrival times for the queueing model.

However, the implications are even more important for *service times* in multiple-server queueing models than for interarrival times. For example, consider the situation where all the servers have the same exponential service-time distribution with parameter μ. For this case, let n be the number of servers *currently* providing service, and let T_i be the *remaining* service time for server i ($i = 1, 2, \ldots, n$), which also has an exponential distribution with parameter $\alpha_i = \mu$. It then follows that U, the time until the *next* service completion from any of these servers, has an exponential distribution with parameter $\alpha = n\mu$. In effect, the queueing system *currently* is performing just like a *single*-server system where service times have an exponential distribution with parameter $n\mu$. We shall make frequent use of this implication for analyzing multiple-server models later in the chapter.

When using this property, it sometimes is useful to also determine the probabilities for *which* of the exponential random variables will turn out to be the one which has the minimum value. For example, you might want to find the probability that a particular server j will finish serving a customer first among n busy exponential servers. It is fairly straightforward (see Prob. 17.4-10) to show that this probability is proportional to the parameter α_j. In particular, the probability that T_j will turn out to be the smallest of the n random variables is

$$P\{T_j = U\} = \alpha_j / \sum_{i=1}^{n} \alpha_i, \qquad \text{for } j = 1, 2, \ldots, n.$$

Property 4: Relationship to the Poisson distribution.

Suppose that the *time* between consecutive occurrences of some particular kind of event (e.g., arrivals or service completions by a continuously busy server) has an exponential distribution with parameter α. Property 4 then has to do with the resulting implication about the probability distribution of the *number* of times this kind of event occurs over a specified time. In particular, let $X(t)$ be the number of occurrences by time t ($t \geq 0$), where time 0 designates the instant at which the count begins. The implication is that

$$P\{X(t) = n\} = \frac{(\alpha t)^n e^{-\alpha t}}{n!}, \qquad \text{for } n = 0, 1, 2, \ldots;$$

that is, $X(t)$ has a Poisson distribution with parameter αt. For example, with $n = 0$,

$$P\{X(t) = 0\} = e^{-\alpha t},$$

which is just the probability from the exponential distribution that the *first* event occurs after time t. The mean of this Poisson distribution is

$$E\{X(t)\} = \alpha t,$$

so that the expected number of events *per unit time* is α. Thus, α is said to be the *mean rate* at which the events occur. When the events are counted on a continuing basis, the counting process $\{X(t); t \geq 0\}$ is said to be a **Poisson process** with parameter α (the mean rate).

This property provides useful information about *service completions* when service times have an exponential distribution with parameter μ. We obtain this information by defining $X(t)$ as the number of service completions achieved by a *continuously busy* server in elapsed time t, where $\alpha = \mu$. For *multiple*-server queueing models, $X(t)$ can also be defined as the number of service completions achieved by n continuously busy servers in elapsed time t, where $\alpha = n\mu$.

The property is particularly useful for describing the probabilistic behavior of *arrivals* when interarrival times have an exponential distribution with parameter λ. In this case, $X(t)$ is the *number* of arrivals in elapsed time t, where $\alpha = \lambda$ is the *mean arrival rate.* Therefore, arrivals occur according to a **Poisson input process** with parameter λ. Such queueing models also are described as assuming a *Poisson input.*

Arrivals sometimes are said to occur *randomly,* meaning that they occur in accordance with a Poisson input process. One intuitive interpretation of this phenomenon is that every time period of fixed length has the *same* chance of having an arrival regardless of when the preceding arrival occurred, as suggested by the following property.

Property 5: For all positive values of t, $P\{T \leq t + \Delta t \mid T > t\} \approx \alpha\, \Delta t$, for small Δt.

Continuing to interpret T as the time from the last event of a certain type (arrival or service completion) until the next such event, we suppose that a time t already has elapsed without the event's occurring. We know from Property 2 that the probability that the event will occur within the next time interval of fixed length Δt is a *constant* (identified in the next paragraph), regardless of how large or small t is. Property 5 goes further to say that when the value of Δt is small, this constant probability can be approximated very closely by $\alpha\, \Delta t$. Furthermore, when considering different small values of Δt, this probability is essentially *proportional* to Δt, with proportionality factor α. In fact, α is the *mean rate* at which the events occur (see Property 4), so that the *expected number* of events in the interval of length Δt is *exactly* $\alpha\, \Delta t$. The only reason that the probability of an event's occurring differs slightly from this value is the possibility that *more than one* event will occur, which has negligible probability when Δt is small.

To see why Property 5 holds mathematically, note that the constant value of our probability (for a fixed value of $\Delta t > 0$) is just

$$\begin{aligned} P\{T \leq t + \Delta t \mid T > t\} &= P\{T \leq \Delta t\} \\ &= 1 - e^{-\alpha\, \Delta t}, \end{aligned}$$

for any $t \geq 0$. Therefore, because the series expansion of e^x for any exponent x is

$$e^x = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!},$$

it follows that

$$P\{T \le t + \Delta t \mid T > t\} = 1 - 1 + \alpha\,\Delta t - \sum_{n=2}^{\infty} \frac{(-\alpha\,\Delta t)^n}{n!}$$
$$\approx \alpha\,\Delta t, \qquad \text{for small } \Delta t,^{1}$$

because the summation terms become relatively negligible for sufficiently small values of $\alpha\,\Delta t$.

Because T can represent either interarrival or service times in queueing models, this property provides a convenient approximation of the probability that the event of interest occurs in the next small interval (Δt) of time. An analysis based on this approximation also can be made exact by taking appropriate limits as $\Delta t \to 0$.

Property 6: Unaffected by aggregation or disaggregation.

This property is relevant primarily for verifying that the *input process* is *Poisson.* Therefore, we shall describe it in these terms, although it also applies directly to the exponential distribution (exponential interarrival times) because of Property 4.

We first consider the aggregation (combining) of several Poisson input processes into one overall input process. In particular, suppose that there are several (n) *different* types of customers, where the customers of each type (type i) arrive according to a *Poisson input process* with parameter λ_i ($i = 1, 2, \ldots, n$). Assuming that these are *independent* Poisson processes, the property says that the *aggregate* input process (arrival of all customers without regard to type) also must be Poisson, with parameter (arrival rate) $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. In other words, having a Poisson process is *unaffected by aggregation.*

This part of the property follows directly from Properties 3 and 4. The latter property implies that the interarrival times for customers of type i have an exponential distribution with parameter λ_i. For this identical situation, we already discussed for Property 3 that it implies that the interarrival times for all customers also must have an exponential distribution, with parameter $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. Using Property 4 again then implies that the aggregate input process is Poisson.

The second part of Property 6 ("unaffected by disaggregation") refers to the reverse case, where the *aggregate* input process (the one obtained by combining the input processes for several customer types) is known to be Poisson with parameter λ, but the question now concerns the nature of the *disaggregated* input processes (the individual input processes for the individual customer types). Assuming that each arriving customer has a *fixed* probability p_i of being of type i ($i = 1, 2, \ldots, n$), with

$$\lambda_i = p_i\lambda \qquad \text{and} \qquad \sum_{i=1}^{n} p_i = 1,$$

[1]More precisely,

$$\lim_{\Delta t \to 0} \frac{P\{T \le t + \Delta t \mid T > t\}}{\Delta t} = \alpha.$$

the property says that the input process for customers of type i also must be Poisson with parameter λ_i. In other words, having a Poisson process is *unaffected by disaggregation.*

As one example of the usefulness of this second part of the property, consider the following situation. Indistinguishable customers arrive according to a Poisson process with parameter λ. Each arriving customer has a fixed probability p of *balking* (leaving without entering the queueing system), so the probability of entering the system is $1 - p$. Thus, there are two types of customers—those who balk and those who enter the system. The property says that each type arrives according to a Poisson process, with parameters $p\lambda$ and $(1 - p)\lambda$, respectively. Therefore, by using the latter Poisson process, queueing models that assume a Poisson input process can still be used to analyze the performance of the queueing system for those customers who enter the system.

17.5 THE BIRTH-AND-DEATH PROCESS

Most elementary queueing models assume that the inputs (arriving customers) and outputs (leaving customers) of the queueing system occur according to the *birth-and-death process.* This important process in probability theory has applications in various areas. However, in the context of queueing theory, the term **birth** refers to the *arrival* of a new customer into the queueing system, and **death** refers to the *departure* of a served customer. The *state* of the system at time t ($t \geq 0$), denoted by $N(t)$, is the number of customers in the queueing system at time t. The birth-and-death process describes *probabilistically* how $N(t)$ changes as t increases. Broadly speaking, it says that *individual* births and deaths occur *randomly,* where their mean occurrence rates depend only upon the current state of the system. More precisely, the assumptions of the birth-and-death process are the following:

Assumption 1. Given $N(t) = n$, the current probability distribution of the *remaining* time until the next *birth* (arrival) is *exponential* with parameter λ_n ($n = 0, 1, 2, \ldots$).

Assumption 2. Given $N(t) = n$, the current probability distribution of the *remaining* time until the next *death* (service completion) is *exponential* with parameter μ_n ($n = 1, 2, \ldots$).

Assumption 3. The random variable of assumption 1 (the remaining time until the next *birth*) and the random variable of assumption 2 (the remaining time until the next *death*) are mutually independent. The next transition in the state of the process is either

$$n \rightarrow n + 1 \quad \text{(a single birth)}$$

or

$$n \rightarrow n - 1 \quad \text{(a single death),}$$

depending on whether the former or latter random variable is smaller.

Because of these assumptions, the birth-and-death process is a special type of *continuous time Markov chain.* (See Sec. 16.8 for a description of continuous time Markov chains and their properties, including an introduction to the general procedure for finding steady-state probabilities that will be applied in the remainder of this section.) Queueing models that can be represented by a continuous time Markov chain are far more tractable analytically than any other.

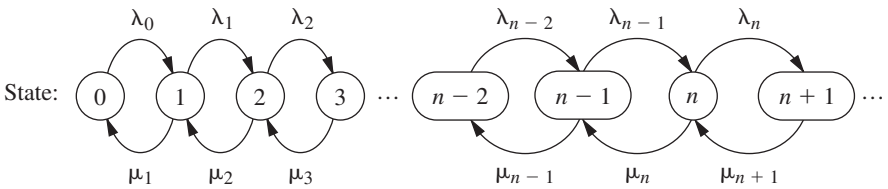


FIGURE 17.4
Rate diagram for the birth-and-death process.

Because Property 4 for the exponential distribution (see Sec. 17.4) implies that the λ_n and μ_n are mean rates, we can summarize these assumptions by the rate diagram shown in Fig. 17.4. The arrows in this diagram show the only possible *transitions* in the state of the system (as specified by assumption 3), and the entry for each arrow gives the mean rate for that transition (as specified by assumptions 1 and 2) when the system is in the state at the base of the arrow.

Except for a few special cases, analysis of the birth-and-death process is very difficult when the system is in a *transient* condition. Some results about the probability distribution of $N(t)$ have been obtained,[1] but they are too complicated to be of much practical use. On the other hand, it is relatively straightforward to derive this distribution *after* the system has reached a *steady-state* condition (assuming that this condition can be reached). This derivation can be done directly from the rate diagram, as outlined next.

Consider any particular state of the system n ($n = 0, 1, 2, \ldots$). Starting at time 0, suppose that a count is made of the number of times that the process enters this state and the number of times it leaves this state, as denoted below:

$E_n(t)$ = number of times that process enters state n by time t.

$L_n(t)$ = number of times that process leaves state n by time t.

Because the two types of events (entering and leaving) must alternate, these two numbers must always either be equal or differ by just 1; that is,

$$|E_n(t) - L_n(t)| \leq 1.$$

Dividing through both sides by t and then letting $t \to \infty$ gives

$$\left|\frac{E_n(t)}{t} - \frac{L_n(t)}{t}\right| \leq \frac{1}{t}, \qquad \text{so} \qquad \lim_{t\to\infty}\left|\frac{E_n(t)}{t} - \frac{L_n(t)}{t}\right| = 0.$$

Dividing $E_n(t)$ and $L_n(t)$ by t gives the *actual rate* (number of events per unit time) at which these two kinds of events have occurred, and letting $t \to \infty$ then gives the *mean rate* (expected number of events per unit time):

$$\lim_{t\to\infty} \frac{E_n(t)}{t} = \text{mean rate at which process enters state } n.$$

$$\lim_{t\to\infty} \frac{L_n(t)}{t} = \text{mean rate at which process leaves state } n.$$

These results yield the following key principle:

[1]S. Karlin and J. McGregor, "Many Server Queueing Processes with Poisson Input and Exponential Service Times," *Pacific Journal of Mathematics,* **8:** 87–118, 1958.

Rate In = Rate Out Principle. For any state of the system n ($n = 0, 1, 2, \ldots$), mean entering rate = mean leaving rate.

The equation expressing this principle is called the **balance equation** for state n. After constructing the balance equations for all the states in terms of the *unknown* P_n probabilities, we can solve this system of equations (plus an equation stating that the probabilities must sum to 1) to find these probabilities.

To illustrate a balance equation, consider state 0. The process enter this state *only* from state 1. Thus, the steady-state probability of being in state 1 (P_1) represents the proportion of time that it would be *possible* for the process to enter state 0. Given that the process is in state 1, the mean rate of entering state 0 is μ_1. (In other words, for each cumulative unit of time that the process spends in state 1, the expected number of times that it would leave state 1 to enter state 0 is μ_1.) From any *other* state, this mean rate is 0. Therefore, the overall mean rate at which the process leaves its current state to enter state 0 (the *mean entering rate*) is

$$\mu_1 P_1 + 0(1 - P_1) = \mu_1 P_1.$$

By the same reasoning, the *mean leaving rate* must be $\lambda_0 P_0$, so the balance equation for state 0 is

$$\mu_1 P_1 = \lambda_0 P_0.$$

For every other state there are two possible transitions both into and out of the state. Therefore, each side of the balance equations for these states represents the *sum* of the mean rates for the two transitions involved. Otherwise, the reasoning is just the same as for state 0. These balance equations are summarized in Table 17.1.

Notice that the first balance equation contains two variables for which to solve (P_0 and P_1), the first two equations contain three variables (P_0, P_1, and P_2), and so on, so that there always is one "extra" variable. Therefore, the procedure in solving these equations is to solve in terms of one of the variables, the most convenient one being P_0. Thus, the first equation is used to solve for P_1 in terms of P_0; this result and the second equation are then used to solve for P_2 in terms of P_0; and so forth. At the end, the requirement that the sum of all the probabilities equal 1 can be used to evaluate P_0.

TABLE 17.1 Balance equations for the birth-and-death process

State	Rate In = Rate Out
0	$\mu_1 P_1 = \lambda_0 P_0$
1	$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$
2	$\lambda_1 P_1 + \mu_3 P_3 = (\lambda_2 + \mu_2) P_2$
$\vdots$	$\vdots$
$n - 1$	$\lambda_{n-2} P_{n-2} + \mu_n P_n = (\lambda_{n-1} + \mu_{n-1}) P_{n-1}$
n	$\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_n) P_n$
$\vdots$	$\vdots$

Applying this procedure yields the following results:

State:

$$
\begin{array}{lllll}
0: & P_1 & = \dfrac{\lambda_0}{\mu_1}P_0 & & \\
1: & P_2 & = \dfrac{\lambda_1}{\mu_2}P_1 + \dfrac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) & = \dfrac{\lambda_1}{\mu_2}P_1 & = \dfrac{\lambda_1\lambda_0}{\mu_2\mu_1}P_0 \\
2: & P_3 & = \dfrac{\lambda_2}{\mu_3}P_2 + \dfrac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) & = \dfrac{\lambda_2}{\mu_3}P_2 & = \dfrac{\lambda_2\lambda_1\lambda_0}{\mu_3\mu_2\mu_1}P_0 \\
\vdots & & \vdots & & \\
n-1: & P_n & = \dfrac{\lambda_{n-1}}{\mu_n}P_{n-1} + \dfrac{1}{\mu_n}(\mu_{n-1}P_{n-1} - \lambda_{n-2}P_{n-2}) & = \dfrac{\lambda_{n-1}}{\mu_n}P_{n-1} & = \dfrac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}P_0 \\
n: & P_{n+1} & = \dfrac{\lambda_n}{\mu_{n+1}}P_n + \dfrac{1}{\mu_{n+1}}(\mu_n P_n - \lambda_{n-1}P_{n-1}) & = \dfrac{\lambda_n}{\mu_{n+1}}P_n & = \dfrac{\lambda_n\lambda_{n-1}\cdots\lambda_0}{\mu_{n+1}\mu_n\cdots\mu_1}P_0 \\
\vdots & & \vdots & &
\end{array}
$$

To simplify notation, let

$$C_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}, \qquad \text{for } n = 1, 2, \ldots,$$

and then define $C_n = 1$ for $n = 0$. Thus, the steady-state probabilities are

$$P_n = C_n P_0, \qquad \text{for } n = 0, 1, 2, \ldots.$$

The requirement that

$$\sum_{n=0}^{\infty} P_n = 1$$

implies that

$$\left(\sum_{n=0}^{\infty} C_n\right)P_0 = 1,$$

so that

$$P_0 = \left(\sum_{n=0}^{\infty} C_n\right)^{-1}.$$

When a queueing model is based on the birth-and-death process, so the state of the system n represents the number of customers in the queueing system, the key measures of performance for the queueing system (L, L_q, W, and W_q) can be obtained immediately

after calculating the P_n from the above formulas. The definitions of L and L_q given in Sec. 17.2 specify that

$$L = \sum_{n=0}^{\infty} nP_n, \qquad L_q = \sum_{n=s}^{\infty} (n - s)P_n.$$

Furthermore, the relationships given at the end of Sec. 17.2 yield

$$W = \frac{L}{\bar{\lambda}}, \qquad W_q = \frac{L_q}{\bar{\lambda}},$$

where $\bar{\lambda}$ is the *average* arrival rate over the long run. Because λ_n is the mean arrival rate while the system is in state n ($n = 0, 1, 2, \ldots$) and P_n is the proportion of time that the system is in this state,

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n.$$

Several of the expressions just given involve summations with an infinite number of terms. Fortunately, these summations have analytic solutions for a number of interesting special cases,[1] as seen in the next section. Otherwise, they can be approximated by summing a finite number of terms on a computer.

These steady-state results have been derived under the assumption that the λ_n and μ_n parameters have values such that the process actually can *reach* a steady-state condition. This assumption *always* holds if $\lambda_n = 0$ for some value of n greater than the initial state, so that only a finite number of states (those less than this n) are possible. It also *always* holds when λ and μ are defined (see "Terminology and Notation" in Sec. 17.2) and $\rho = \lambda/(s\mu) < 1$. It does *not* hold if $\Sigma_{n=1}^{\infty} C_n = \infty$.

The following section describes several queueing models that are special cases of the birth-and-death process. Therefore, the general steady-state results just given in boxes will be used over and over again to obtain the specific steady-state results for these models.

17.6 QUEUEING MODELS BASED ON THE BIRTH-AND-DEATH PROCESS

Because each of the mean rates $\lambda_0, \lambda_1, \ldots$ and $\mu_1, \mu_2, \ldots$ for the birth-and-death process can be assigned any nonnegative value, we have great flexibility in modeling a queueing system. Probably the most widely used models in queueing theory are based directly upon

[1]These solutions are based on the following known results for the sum of any geometric series:

$$\sum_{n=0}^{N} x^n = \frac{1 - x^{N+1}}{1 - x}, \qquad \text{for any } x \neq 1,$$

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1 - x}, \qquad \text{if } |x| < 1.$$

this process. Because of assumptions 1 and 2 (and Property 4 for the exponential distribution), these models are said to have a **Poisson input** and **exponential service times.** The models differ only in their assumptions about how the λ_n and μ_n change with n. We present four of these models in this section for four important types of queueing systems.

The *M/M/s* Model

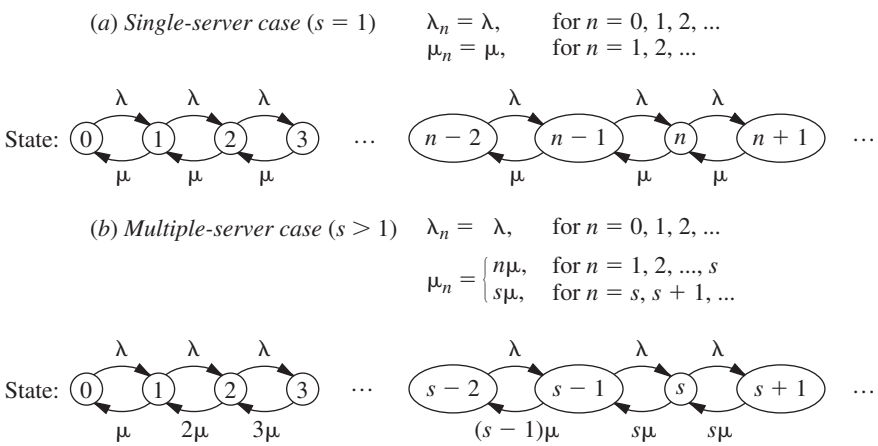
As described in Sec. 17.2, the *M/M/s* model assumes that all *interarrival times* are independently and identically distributed according to an exponential distribution (i.e., the input process is Poisson), that all *service times* are independent and identically distributed according to another exponential distribution, and that the number of servers is s (any positive integer). Consequently, this model is just the special case of the birth-and-death process where the queueing system's *mean arrival rate* and *mean service rate per busy server* are constant (λ and μ, respectively) regardless of the state of the system. When the system has just a *single server* ($s = 1$), the implication is that the parameters for the birth-and-death process are $\lambda_n = \lambda$ ($n = 0, 1, 2, \ldots$) and $\mu_n = \mu$ ($n = 1, 2, \ldots$). The resulting rate diagram is shown in Fig. 17.5*a*.

However, when the system has *multiple servers* ($s > 1$), the μ_n cannot be expressed this simply. Keep in mind that μ_n represents the mean service rate for the *overall* queueing system (i.e., the mean rate at which service completions occur, so that customers leave the system) when there are n customers currently in the system. As mentioned for Property 4 of the exponential distribution (see Sec. 17.4), when the mean service rate per busy server is μ, the overall mean service rate for n busy servers must be $n\mu$. Therefore, $\mu_n = n\mu$ when $n \leq s$, whereas $\mu_n = s\mu$ when $n \geq s$ so that all s servers are busy. The rate diagram for this case is shown in Fig. 17.5*b*.

When the maximum mean service rate $s\mu$ exceeds the mean arrival rate λ, that is, when

$$\rho = \frac{\lambda}{s\mu} < 1,$$

FIGURE 17.5
Rate diagrams for the *M/M/s* model.



a queueing system fitting this model will eventually reach a steady-state condition. In this situation, the steady-state results derived in Sec. 17.5 for the general birth-and-death process are directly applicable. However, these results simplify considerably for this model and yield closed-form expressions for P_n, L, L_q, and so forth, as shown next.

Results for the Single-Server Case (*M*/*M*/1). For $s = 1$, the C_n factors for the birth-and-death process reduce to

$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n, \qquad \text{for } n = 0, 1, 2, \ldots$$

Therefore,

$$P_n = \rho^n P_0, \qquad \text{for } n = 0, 1, 2, \ldots,$$

where

$$\begin{aligned} P_0 &= \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1} \\ &= \left(\frac{1}{1-\rho}\right)^{-1} \\ &= 1 - \rho. \end{aligned}$$

Thus,

$$P_n = (1-\rho)\rho^n, \qquad \text{for } n = 0, 1, 2, \ldots.$$

Consequently,

$$\begin{aligned} L &= \sum_{n=0}^{\infty} n(1-\rho)\rho^n \\ &= (1-\rho)\rho \sum_{n=0}^{\infty} \frac{d}{d\rho}(\rho^n) \\ &= (1-\rho)\rho \frac{d}{d\rho}\left(\sum_{n=0}^{\infty} \rho^n\right) \\ &= (1-\rho)\rho \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) \\ &= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}. \end{aligned}$$

Similarly,

$$\begin{aligned} L_q &= \sum_{n=1}^{\infty} (n-1)P_n \\ &= L - 1(1 - P_0) \\ &= \frac{\lambda^2}{\mu(\mu - \lambda)}. \end{aligned}$$

When $\lambda \geq \mu$, so that the mean arrival rate exceeds the mean service rate, the preceding solution "blows up" (because the summation for computing P_0 diverges). For this case, the queue would "explode" and grow without bound. If the queueing system begins operation with no customers present, the server might succeed in keeping up with arriving customers over a short period of time, but this is impossible in the long run. (Even when $\lambda = \mu$, the *expected* number of customers in the queueing system slowly grows without bound over time because, even though a temporary return to no customers present always is possible, the probabilities of huge numbers of customers present become increasingly significant over time.)

Assuming again that $\lambda < \mu$, we now can derive the probability distribution of the *waiting time in the system* (so *including* service time) $\mathcal{W}$ for a random arrival when the queue discipline is first-come-first-served. If this arrival finds n customers already in the system, then the arrival will have to wait through $n + 1$ exponential service times, including his or her own. (For the customer currently being served, recall the lack-of-memory property for the exponential distribution discussed in Sec. 17.4.) Therefore, let $T_1, T_2, \ldots$ be independent service-time random variables having an exponential distribution with parameter μ, and let

$$S_{n+1} = T_1 + T_2 + \cdots + T_{n+1}, \qquad \text{for } n = 0, 1, 2, \ldots,$$

so that S_{n+1} represents the *conditional* waiting time given n customers already in the system. As discussed in Sec. 17.7, S_{n+1} is known to have an *Erlang distribution.*[1] Because the probability that the random arrival will find n customers in the system is P_n, it follows that

$$P\{\mathcal{W} > t\} = \sum_{n=0}^{\infty} P_n P\{S_{n+1} > t\},$$

which reduces after considerable manipulation (see Prob. 17.6-17) to

$$P\{\mathcal{W} > t\} = e^{-\mu(1-\rho)t}, \qquad \text{for } t \geq 0.$$

The surprising conclusion is that $\mathcal{W}$ has an *exponential* distribution with parameter $\mu(1 - \rho)$. Therefore,

$$W = E(\mathcal{W}) = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu - \lambda}.$$

These results *include* service time in the waiting time. In some contexts (e.g., the County Hospital emergency room problem), the more relevant waiting time is just until service begins. Thus, consider the *waiting time in the queue* (so *excluding* service time) $\mathcal{W}_q$ for a random arrival when the queue discipline is first-come-first-served. If this arrival finds no customers already in the system, then the arrival is served immediately, so that

$$P\{\mathcal{W}_q = 0\} = P_0 = 1 - \rho.$$

[1]Outside queueing theory, this distribution is known as the *gamma distribution.*

If this arrival finds $n > 0$ customers already there instead, then the arrival has to wait through n exponential service times until his or her own service begins, so that

$$\begin{aligned} P\{\mathcal{W}_q > t\} &= \sum_{n=1}^{\infty} P_n P\{S_n > t\} \\ &= \sum_{n=1}^{\infty} (1-\rho)\rho^n P\{S_n > t\} \\ &= \rho \sum_{n=0}^{\infty} P_n P\{S_{n+1} > t\} \\ &= \rho P\{\mathcal{W} > t\} \\ &= \rho e^{-\mu(1-\rho)t}, \qquad \text{for } t \geq 0. \end{aligned}$$

Note that Wq does not quite have an exponential distribution, because $P\{\mathcal{W}_q = 0\} > 0$. However, the *conditional* distribution of $\mathcal{W}_q$, given that $\mathcal{W}_q > 0$, does have an exponential distribution with parameter $\mu(1-\rho)$, just as $\mathcal{W}$ does, because

$$P\{\mathcal{W}_q > t \mid \mathcal{W}_q > 0\} = \frac{P\{\mathcal{W}_q > t\}}{P\{\mathcal{W}_q > 0\}} = e^{-\mu(1-\rho)t}, \qquad \text{for } t \geq 0.$$

By deriving the mean of the (unconditional) distribution of $\mathcal{W}_q$ (or applying either $L_q = \lambda W_q$ or $W_q = W - 1/\mu$),

$$W_q = E(\mathcal{W}_q) = \frac{\lambda}{\mu(\mu - \lambda)}.$$

Results for the Multiple-Server Case ($s > 1$). When $s > 1$, the C_n factors become

$$C_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} & \text{for } n = 1, 2, \ldots, s \\[2ex] \dfrac{(\lambda/\mu)^s}{s!}\left(\dfrac{\lambda}{s\mu}\right)^{n-s} = \dfrac{(\lambda/\mu)^n}{s!\,s^{n-s}} & \text{for } n = s, s+1, \ldots. \end{cases}$$

Consequently, if $\lambda < s\mu$ [so that $\rho = \lambda/(s\mu) < 1$], then

$$\begin{aligned} P_0 &= 1 \Bigg/ \left[1 + \sum_{n=1}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s}^{\infty} \left(\frac{\lambda}{s\mu}\right)^{n-s}\right] \\ &= 1 \Bigg/ \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1 - \lambda/(s\mu)}\right], \end{aligned}$$

where the $n = 0$ term in the last summation yields the correct value of 1 because of the convention that $n! = 1$ when $n = 0$. These C_n factors also give

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & \text{if } 0 \leq n \leq s \\[2ex] \dfrac{(\lambda/\mu)^n}{s!\,s^{n-s}} P_0 & \text{if } n \geq s. \end{cases}$$

Furthermore,

$$
\begin{aligned}
L_q &= \sum_{n=s}^{\infty} (n-s)P_n \\
&= \sum_{j=0}^{\infty} jP_{s+j} \\
&= \sum_{j=0}^{\infty} j\,\frac{(\lambda/\mu)^s}{s!}\rho^j P_0 \\
&= P_0\frac{(\lambda/\mu)^s}{s!}\rho \sum_{j=0}^{\infty} \frac{d}{d\rho}(\rho^j) \\
&= P_0\frac{(\lambda/\mu)^s}{s!}\rho \frac{d}{d\rho}\left(\sum_{j=0}^{\infty} \rho^j\right) \\
&= P_0\frac{(\lambda/\mu)^s}{s!}\rho \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) \\
&= \frac{P_0(\lambda/\mu)^s\rho}{s!(1-\rho)^2}; \\
W_q &= \frac{L_q}{\lambda}; \\
W &= W_q + \frac{1}{\mu}; \\
L &= \lambda\left(W_q + \frac{1}{\mu}\right) = L_q + \frac{\lambda}{\mu}.
\end{aligned}
$$

Figures 17.6 and 17.7 show how P_0 and L change with ρ for various values of s.

The single-server method for finding the probability distribution of waiting times also can be extended to the multiple-server case. This yields[1] (for $t \geq 0$)

$$P\{\mathcal{W} > t\} = e^{-\mu t}\left[\frac{1 + P_0(\lambda/\mu)^s}{s!(1-\rho)}\left(\frac{1 - e^{-\mu t(s-1-\lambda/\mu)}}{s - 1 - \lambda/\mu}\right)\right]$$

and

$$P\{\mathcal{W}_q > t\} = (1 - P\{\mathcal{W}_q = 0\})e^{-s\mu(1-\rho)t},$$

where

$$P\{\mathcal{W}_q = 0\} = \sum_{n=0}^{s-1} P_n.$$

The above formulas for the various measures of performance (including the P_n) are relatively imposing for hand calculations. However, this chapter's Excel file in your OR

[1]When $s - 1 - \lambda/\mu = 0$, $(1 - e^{-\mu t(s-1-\lambda/\mu)})/(s - 1 - \lambda/\mu)$ should be replaced by μt.

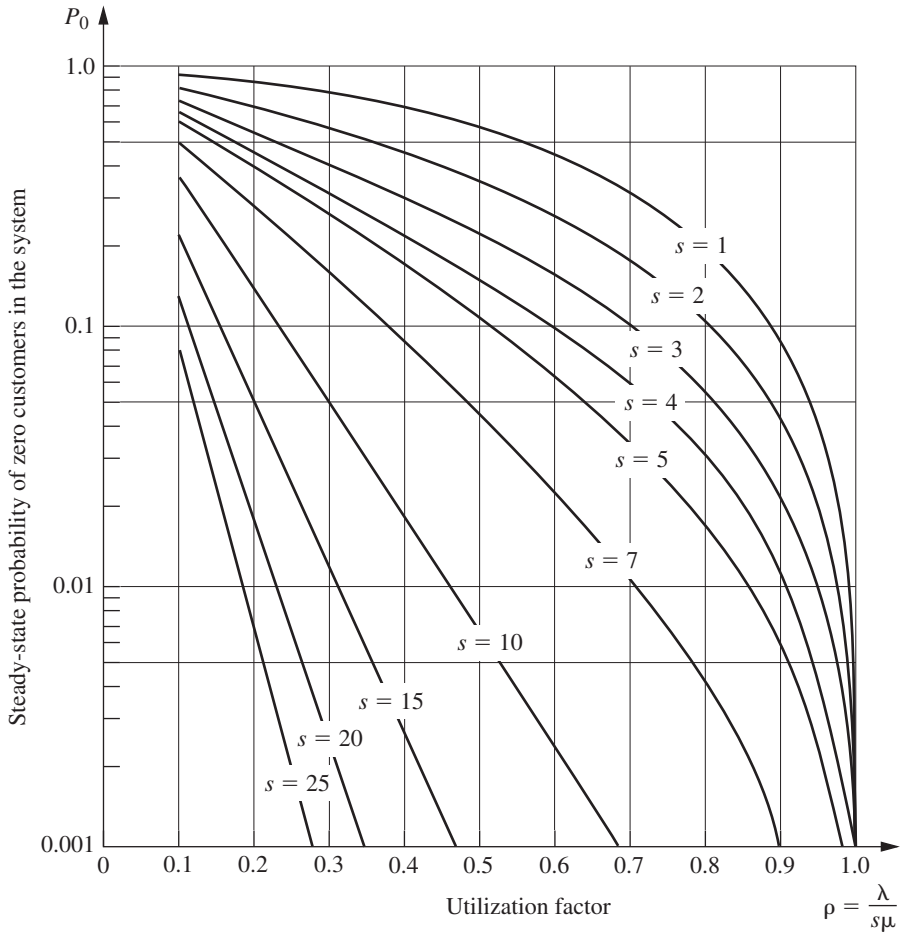


FIGURE 17.6
Values of P_0 for the $M/M/s$ model (Sec. 17.6).

Courseware includes an Excel template that performs all these calculations simultaneously for any values of t, s, λ, and μ you want, provided that $\lambda < s\mu$.

If $\lambda \geq s\mu$, so that the mean arrival rate exceeds the maximum mean service rate, then the queue grows without bound, so the preceding steady-state solutions are not applicable.

The County Hospital Example with the $M/M/s$ Model. For the County Hospital emergency room problem (see Sec. 17.1), the management engineer has concluded that the emergency cases arrive pretty much at random (a *Poisson input process*), so that interarrival times have an exponential distribution. She also has concluded that the time spent by a doctor treating the cases approximately follows an *exponential distribution.* Therefore, she has chosen the $M/M/s$ model for a preliminary study of this queueing system.

By projecting the available data for the early evening shift into next year, she estimates that patients will arrive at an *average* rate of 1 every $\frac{1}{2}$ hour. A doctor re-

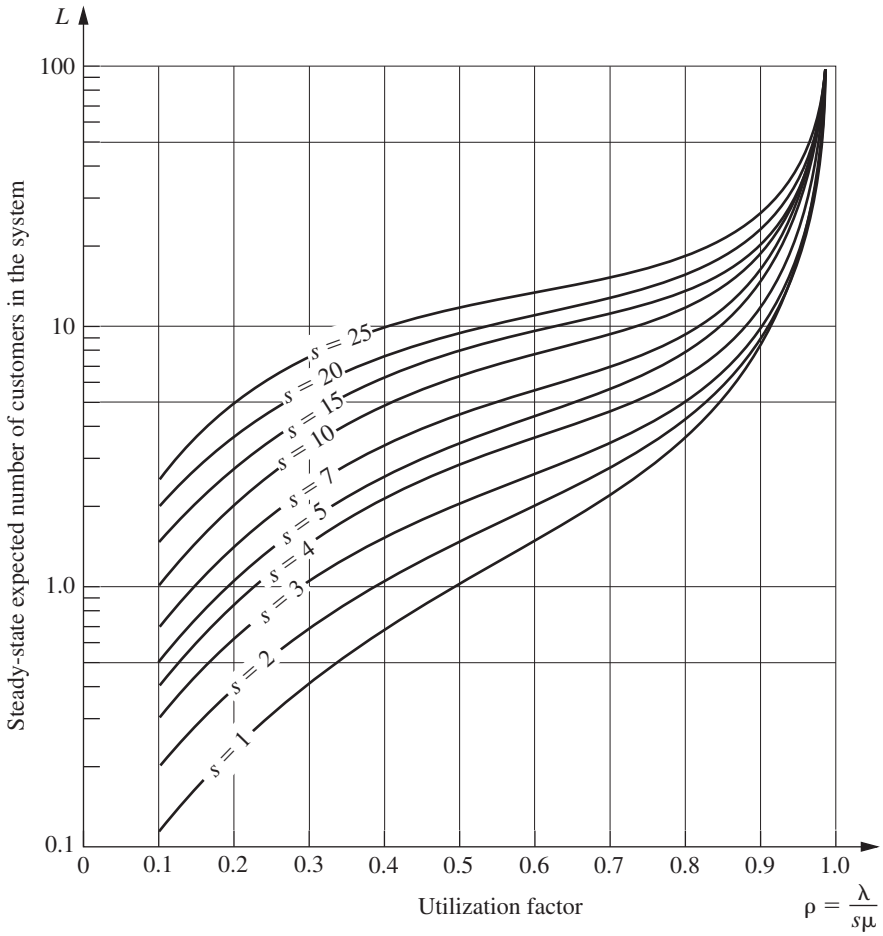


FIGURE 17.7
Values for L for the $M/M/s$ model (Sec. 17.6).

quires an average of 20 minutes to treat each patient. Thus, with one hour as the unit of time,

$$\frac{1}{\lambda} = \frac{1}{2} \text{ hour per customer}$$

and

$$\frac{1}{\mu} = \frac{1}{3} \text{ hour per customer,}$$

so that

$$\lambda = 2 \text{ customers per hour}$$

and

$$\mu = 3 \text{ customers per hour.}$$

The two alternatives being considered are to continue having just one doctor during this shift ($s = 1$) or to add a second doctor ($s = 2$). In both cases,

$$\rho = \frac{\lambda}{s\mu} < 1,$$

so that the system should approach a steady-state condition. (Actually, because λ is somewhat different during other shifts, the system will never truly reach a steady-state condition, but the management engineer feels that steady-state results will provide a good approximation.) Therefore, the preceding equations are used to obtain the results shown in Table 17.2.

On the basis of these results, she tentatively concluded that a single doctor would be inadequate next year for providing the relatively prompt treatment needed in a hospital emergency room. You will see later how she checked this conclusion by applying two other queueing models that provide better representations of the real queueing system in some ways.

TABLE 17.2 Steady-state results from the $M/M/s$ model for the County Hospital problem

	$s = 1$	$s = 2$
ρ	$\frac{2}{3}$	$\frac{1}{3}$
P_0	$\frac{1}{3}$	$\frac{1}{2}$
P_1	$\frac{2}{9}$	$\frac{1}{3}$
P_n for $n \geq 2$	$\frac{1}{3}\left(\frac{2}{3}\right)^n$	$\left(\frac{1}{3}\right)^n$
L_q	$\frac{4}{3}$	$\frac{1}{12}$
L	2	$\frac{3}{4}$
W_q	$\frac{2}{3}$ hour	$\frac{1}{24}$ hour
W	1 hour	$\frac{3}{8}$ hour
$P\{\mathcal{W}_q > 0\}$	0.667	0.167
$P\left\{\mathcal{W}_q > \frac{1}{2}\right\}$	0.404	0.022
$P\{\mathcal{W}_q > 1\}$	0.245	0.003
$P\{\mathcal{W}_q > t\}$	$\frac{2}{3}e^{-t}$	$\frac{1}{6}e^{-4t}$
$P\{\mathcal{W} > t\}$	e^{-t}	$\frac{1}{2}e^{-3t}(3 - e^{-t})$

The Finite Queue Variation of the *M/M/s* Model (Called the *M/M/s/K* Model)

We mentioned in the discussion of queues in Sec. 17.2 that queueing systems sometimes have a *finite queue;* i.e., the number of customers in the system is not permitted to exceed some specified number (denoted by K) so the queue capacity is $K - s$. Any customer that arrives while the queue is "full" is refused entry into the system and so leaves forever. From the viewpoint of the birth-and-death process, the mean input rate into the system becomes zero at these times. Therefore, the one modification needed in the $M/M/s$ model to introduce a finite queue is to change the λ_n parameters to

$$\lambda_n = \begin{cases} \lambda & \text{for } n = 0, 1, 2, \ldots, K-1 \\ 0 & \text{for } n \geq K. \end{cases}$$

Because $\lambda_n = 0$ for some values of n, a queueing system that fits this model always will eventually reach a steady-state condition, even when $\rho = \lambda/s\mu \geq 1$.

This model commonly is labeled $M/M/s/K$, where the presence of the fourth symbol distinguishes it from the $M/M/s$ model. The single difference in the formulation of these two models is that K is finite for the $M/M/s/K$ model and $K = \infty$ for the $M/M/s$ model.

The usual physical interpretation for the $M/M/s/K$ model is that there is only *limited waiting room* that will accommodate a maximum of K customers in the system. For example, for the County Hospital emergency room problem, this system actually would have a finite queue if there were only K cots for the patients and if the policy were to send arriving patients to another hospital whenever there were no empty cots.

Another possible interpretation is that arriving customers will leave and "take their business elsewhere" whenever they find too many customers (K) ahead of them in the system because they are not willing to incur a long wait. This balking phenomenon is quite common in commercial service systems. However, there are other models available (e.g., see Prob. 17.5-5) that fit this interpretation even better.

The rate diagram for this model is identical to that shown in Fig. 17.5 for the $M/M/s$ model, *except* that it stops with state K.

Results for the Single-Server Case ($M/M/1/K$). For this case,

$$C_n = \begin{cases} \left(\dfrac{\lambda}{\mu}\right)^n = \rho^n & \text{for } n = 0, 1, 2, \ldots, K \\ 0 & \text{for } n > K. \end{cases}$$

Therefore, for $\rho \neq 1$,[1]

$$\begin{aligned} P_0 &= \frac{1}{\sum_{n=0}^{K} (\lambda/\mu)^n} \\ &= 1 \Bigg/ \left[\frac{1 - (\lambda/\mu)^{K+1}}{1 - \lambda/\mu}\right] \\ &= \frac{1 - \rho}{1 - \rho^{K+1}}, \end{aligned}$$

[1]If $\rho = 1$, then $P_n = 1/(K + 1)$ for $n = 0, 1, 2, \ldots, K$, so that $L = K/2$.

so that

$$P_n = \frac{1-\rho}{1-\rho^{K+1}}\rho^n, \qquad \text{for } n = 0, 1, 2, \ldots, K.$$

Hence,

$$\begin{aligned}
L &= \sum_{n=0}^{K} nP_n \\
&= \frac{1-\rho}{1-\rho^{K+1}}\rho \sum_{n=0}^{K} \frac{d}{d\rho}(\rho^n) \\
&= \frac{1-\rho}{1-\rho^{K+1}}\rho \frac{d}{d\rho}\left(\sum_{n=0}^{K} \rho^n\right) \\
&= \frac{1-\rho}{1-\rho^{K+1}}\rho \frac{d}{d\rho}\left(\frac{1-\rho^{K+1}}{1-\rho}\right) \\
&= \rho\frac{-(K+1)\rho^K + K\rho^{K+1} + 1}{(1-\rho^{K+1})(1-\rho)} \\
&= \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}.
\end{aligned}$$

As usual (when $s = 1$),

$$L_q = L - (1 - P_0).$$

Notice that the preceding results do not require that $\lambda < \mu$ (i.e., that $\rho < 1$).

When $\rho < 1$, it can be verified that the second term in the final expression for L converges to 0 as $K \to \infty$, so that *all* the preceding results do indeed converge to the corresponding results given earlier for the $M/M/1$ model.

The waiting-time distributions can be derived by using the same reasoning as for the $M/M/1$ model (see Prob. 17.6-31). However, no simple expressions are obtained in this case, so computer calculations are required. Fortunately, even though $L \neq \lambda W$ and $L_q \neq \lambda W_q$ for the current model because the λ_n are not equal for all n (see the end of Sec. 17.2), the *expected* waiting times for customers entering the system still can be obtained directly from the expressions given at the end of Sec. 17.5:

$$W = \frac{L}{\bar{\lambda}}, \qquad W_q = \frac{L_q}{\bar{\lambda}},$$

where

$$\begin{aligned}
\bar{\lambda} &= \sum_{n=0}^{\infty} \lambda_n P_n \\
&= \sum_{n=0}^{K-1} \lambda P_n \\
&= \lambda(1 - P_K).
\end{aligned}$$

Results for the Multiple-Server Case ($s > 1$). Because this model does not allow more than K customers in the system, K is the maximum number of servers that could ever be used. Therefore, assume that $s \leq K$. In this case, C_n becomes

$$C_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} & \text{for } n = 0, 1, 2, \ldots, s \\ \dfrac{(\lambda/\mu)^s}{s!}\left(\dfrac{\lambda}{s\mu}\right)^{n-s} = \dfrac{(\lambda/\mu)^n}{s!\,s^{n-s}} & \text{for } n = s, s+1, \ldots, K \\ 0 & \text{for } n > K. \end{cases}$$

Hence,

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!}P_0 & \text{for } n = 1, 2, \ldots, s \\ \dfrac{(\lambda/\mu)^n}{s!\,s^{n-s}}P_0 & \text{for } n = s, s+1, \ldots, K \\ 0 & \text{for } n > K, \end{cases}$$

where

$$P_0 = 1 \Bigg/ \left[\sum_{n=0}^{s} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\sum_{n=s+1}^{K}\left(\frac{\lambda}{s\mu}\right)^{n-s}\right].$$

Adapting the derivation of L_q for the $M/M/s$ model to this case (see Prob. 17.6-28) yields

$$L_q = \frac{P_0(\lambda/\mu)^s\rho}{s!(1-\rho)^2}\,[1 - \rho^{K-s} - (K-s)\rho^{K-s}(1-\rho)],$$

where $\rho = \lambda/(s\mu)$.[1] It can then be shown (see Prob. 17.2-5) that

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right).$$

And W and W_q are obtained from these quantities just as shown for the single-server case.

This chapter's Excel file includes an Excel template for calculating the above measures of performance (including the P_n) for this model.

One interesting special case of this model is where $K = s$ so the queue capacity is $K - s = 0$. In this case, customers who arrive when all servers are busy will leave immediately and be lost to the system. This would occur, for example, in a telephone network with s trunk lines so callers get a busy signal and hang up when all the trunk lines are busy. This kind of system (a "queueing system" with no queue) is referred to as *Erlang's loss system* because it was first studied in the early 20th century by A. K. Erlang, a Danish telephone engineer who is considered the founder of queueing theory.

[1]If $\rho = 1$, it is necessary to apply L'Hôpital's rule twice to this expression for L_q. Otherwise, all these multiple-server results hold for all $\rho > 0$. The reason that this queueing system can reach a steady-state condition even when $\rho \geq 1$ is that $\lambda_n = 0$ for $n \geq K$, so that the number of customers in the system cannot continue to grow indefinitely.

The Finite Calling Population Variation of the *M/M/s* Model

Now assume that the only deviation from the *M/M/s* model is that (as defined in Sec. 17.2) the *input source* is *limited;* i.e., the size of the *calling population* is *finite.* For this case, let N denote the size of the calling population. Thus, when the number of customers in the queueing system is n ($n = 0, 1, 2, \ldots, N$), there are only $N - n$ *potential* customers remaining in the input source.

The most important application of this model has been to the machine repair problem, where one or more maintenance people are assigned the responsibility of maintaining in operational order a certain group of N machines by repairing each one that breaks down. (The example given at the end of Sec. 16.8 illustrates this application when the general procedures for solving any *continuous time Markov chain* are used rather than the specific formulas available for the birth-and-death process.) The maintenance people are considered to be individual servers in the queueing system if they work individually on different machines, whereas the entire crew is considered to be a single server if crew members work together on each machine. The machines constitute the calling population. Each one is considered to be a customer in the queueing system when it is down waiting to be repaired, whereas it is outside the queueing system while it is operational.

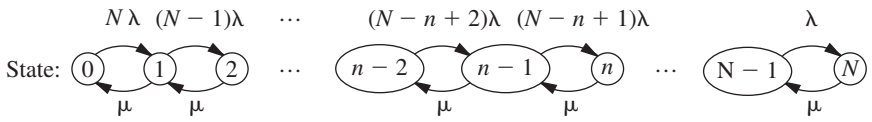
Note that each member of the calling population alternates between being *inside* and *outside* the queueing system. Therefore, the analog of the *M/M/s* model that fits this situation assumes that *each* member's *outside time* (i.e., the elapsed time from leaving the system until returning for the next time) has an *exponential distribution* with parameter λ. When n of the members are *inside,* and so $N - n$ members are *outside,* the current probability distribution of the *remaining* time until the next arrival to the queueing system is the distribution of the *minimum* of the *remaining outside times* for the latter $N - n$ members. Properties 2 and 3 for the exponential distribution imply that this distribution must be exponential with parameter $\lambda_n = (N - n)\lambda$. Hence, this model is just the special case of the birth-and-death process that has the rate diagram shown in Fig. 17.8.

FIGURE 17.8
Rate diagrams for the finite calling population variation of the *M/M/s* model.

(*a*) *Single-server case* ($s = 1$)

$$\lambda_n = \begin{cases} (N-n)\lambda, & \text{for } n = 0, 1, 2, \ldots, N \\ 0, & \text{for } n \geq N \end{cases}$$

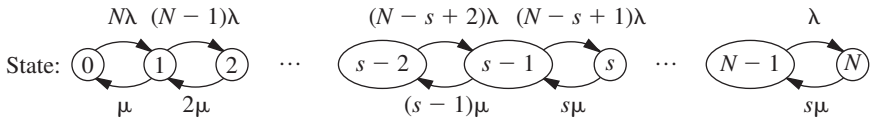
$$\mu_n = \mu, \quad \text{for } n = 1, 2, \ldots$$



(*b*) *Multiple-server case* ($s > 1$)

$$\lambda_n = \begin{cases} (N-n)\lambda, & \text{for } n = 0, 1, 2, \ldots, N \\ 0, & \text{for } n \geq N \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & \text{for } n = 1, 2, \ldots, s \\ s\mu, & \text{for } n = s, s+1, \ldots \end{cases}$$



Because $\lambda_n = 0$ for $n = N$, any queueing system that fits this model will eventually reach a steady-state condition. The available steady-state results are summarized as follows:

Results for the Single-Server Case ($s = 1$). When $s = 1$, the C_n factors in Sec. 17.5 reduce to

$$C_n = \begin{cases} N(N-1)\cdots(N-n+1)\left(\dfrac{\lambda}{\mu}\right)^n = \dfrac{N!}{(N-n)!}\left(\dfrac{\lambda}{\mu}\right)^n & \text{for } n \le N \\ 0 & \text{for } n > N, \end{cases}$$

for this model. Therefore,

$$P_0 = 1\Big/\sum_{n=0}^{N}\left[\frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n\right];$$

$$P_n = \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n P_0, \qquad \text{if } n = 1, 2, \ldots, N;$$

$$L_q = \sum_{n=1}^{N}(n-1)P_n,$$

which can be reduced to

$$L_q = N - \frac{\lambda+\mu}{\lambda}(1 - P_0);$$

$$L = \sum_{n=0}^{N} nP_n = L_q + 1 - P_0$$

$$= N - \frac{\mu}{\lambda}(1 - P_0).$$

Finally,

$$W = \frac{L}{\bar{\lambda}} \qquad \text{and} \qquad W_q = \frac{L_q}{\bar{\lambda}},$$

where

$$\bar{\lambda} = \sum_{n=0}^{\infty}\lambda_n P_n = \sum_{n=0}^{N}(N-n)\lambda P_n = \lambda(N - L).$$

Results for the Multiple-Server Case ($s > 1$). For $N \ge s > 1$,

$$C_n = \begin{cases} \dfrac{N!}{(N-n)!\,n!}\left(\dfrac{\lambda}{\mu}\right)^n & \text{for } n = 0, 1, 2, \ldots, s \\ \dfrac{N!}{(N-n)!\,s!\,s^{n-s}}\left(\dfrac{\lambda}{\mu}\right)^n & \text{for } n = s, s+1, \ldots, N \\ 0 & \text{for } n > N. \end{cases}$$

Hence,

$$P_n = \begin{cases} \dfrac{N!}{(N-n)!\,n!}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & \text{if } 0 \le n \le s \\ \dfrac{N!}{(N-n)!\,s!\,s^{n-s}}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & \text{if } s \le n \le N \\ 0 & \text{if } n > N, \end{cases}$$

where

$$P_0 = 1 \Bigg/ \left[\sum_{n=0}^{s-1} \frac{N!}{(N-n)!\,n!}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=s}^{N} \frac{N!}{(N-n)!\,s!\,s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n \right].$$

Finally,

$$L_q = \sum_{n=s}^{N} (n-s)P_n$$

and

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right),$$

which then yield W and W_q by the same equations as in the single-server case.

This chapter's Excel file includes an Excel template for performing all the above calculations.

Extensive tables of computational results also are available[1] for this model for both the single-server and multiple-server cases.

For both cases, it has been shown[2] that the preceding formulas for P_n and P_0 (and so for L_q, L, W, and W_q) *also* hold for a generalization of this model. In particular, we can *drop* the assumption that the times spent *outside* the queueing system by the members of the calling population have an *exponential distribution,* even though this takes the model outside the realm of the birth-and-death process. As long as these times are identically distributed with mean $1/\lambda$ (and the assumption of exponential service times still holds), these outside times can have *any* probability distribution!

A Model with State-Dependent Service Rate and/or Arrival Rate

All the models thus far have assumed that the mean service rate is always a constant, regardless of how many customers are in the system. Unfortunately, this rate often is not a constant in real queueing systems, particularly when the servers are people. When there is a large backlog of work (i.e., a long queue), it is quite likely that such servers will tend to work faster than they do when the backlog is small or nonexistent. This increase in the service rate may result merely because the servers increase their efforts when they are under the pressure of a long queue. However, it may also result partly because the quality of the service is compromised or because assistance is obtained on certain service phases.

[1]L. G. Peck and R. N. Hazelwood, *Finite Queueing Tables,* Wiley, New York, 1958.

[2]B. D. Bunday and R. E. Scraton, "The G/M/r Machine Interference Model," *European Journal of Operational Research,* **4:** 399–402, 1980.

Given that the mean service rate does increase as the queue size increases, it is desirable to develop a theoretical model that seems to describe the pattern by which it increases. This model not only should be a reasonable approximation of the actual pattern but also should be simple enough to be practical for implementation. One such model is formulated next. (You have the flexibility to formulate many similar models within the framework of the birth-and-death process.) We then show how the same results apply when the arrival rate is affected by the queue size in an analogous way.

Formulation for the Single-Server Case ($s = 1$). Let

$$\mu_n = n^c \mu_1, \qquad \text{for } n = 1, 2, \ldots,$$

where n = number of customers in system,

μ_n = mean service rate when n customers are in system,

$1/\mu_1$ = expected "normal" service time—expected time to service customer when that customer is only one in system,

c = pressure coefficient—positive constant that indicates degree to which service rate of system is affected by system state.

Thus, by selecting $c = 1$, for example, we hypothesize that the mean service rate is directly proportional to n; $c = \frac{1}{2}$ implies that the mean service rate is proportional to the square root of n; and so on. The preceding queueing models in this section have implicitly assumed that $c = 0$.

Now assume additionally that the queueing system has a Poisson input with $\lambda_n = \lambda$ (for $n = 0, 1, 2, \ldots$) and exponential service times with μ_n as just given. This case is now a special case of the birth-and-death process, where

$$C_n = \frac{(\lambda/\mu_1)^n}{(n!)^c}, \qquad \text{for } n = 0, 1, 2, \ldots.$$

Thus, all the steady-state results given in Sec. 17.5 are applicable to this model. (A steady-state condition always can be reached when $c > 0$.) Unfortunately, analytical expressions are not available for the summations involved. However, nearly exact values of P_0 and L have been tabulated[1] for various values of c and λ/μ_1 by summing a finite number of terms on a computer. A small portion of these results also is shown in Figs. 17.9 and 17.10.

A queueing system may react to a long queue by decreasing the arrival rate instead of increasing the service rate. (The arrival rate may be decreased, e.g., by diverting some of the customers requiring service to another service facility.) The corresponding model for describing mean arrival rates for this case lets

$$\lambda_n = (n + 1)^{-b}\lambda_0, \qquad \text{for } n = 0, 1, 2, \ldots,$$

where b is a constant whose interpretation is analogous to that for c. The C_n values for the birth-and-death process with these λ_n (and with $\mu_n = \mu$ for $n = 1, 2, \ldots$) are *identical* to those just shown (replacing λ by λ_0) for the state-dependent service rate model when $c = b$ and $\lambda/\mu_1 = \lambda_0/\mu$, so the steady-state results also are the same.

[1]R. W. Conway and W. L. Maxwell, "A Queueing Model with State Dependent Service Rate," *Journal of Industrial Engineering,* **12:** 132–136, 1961.

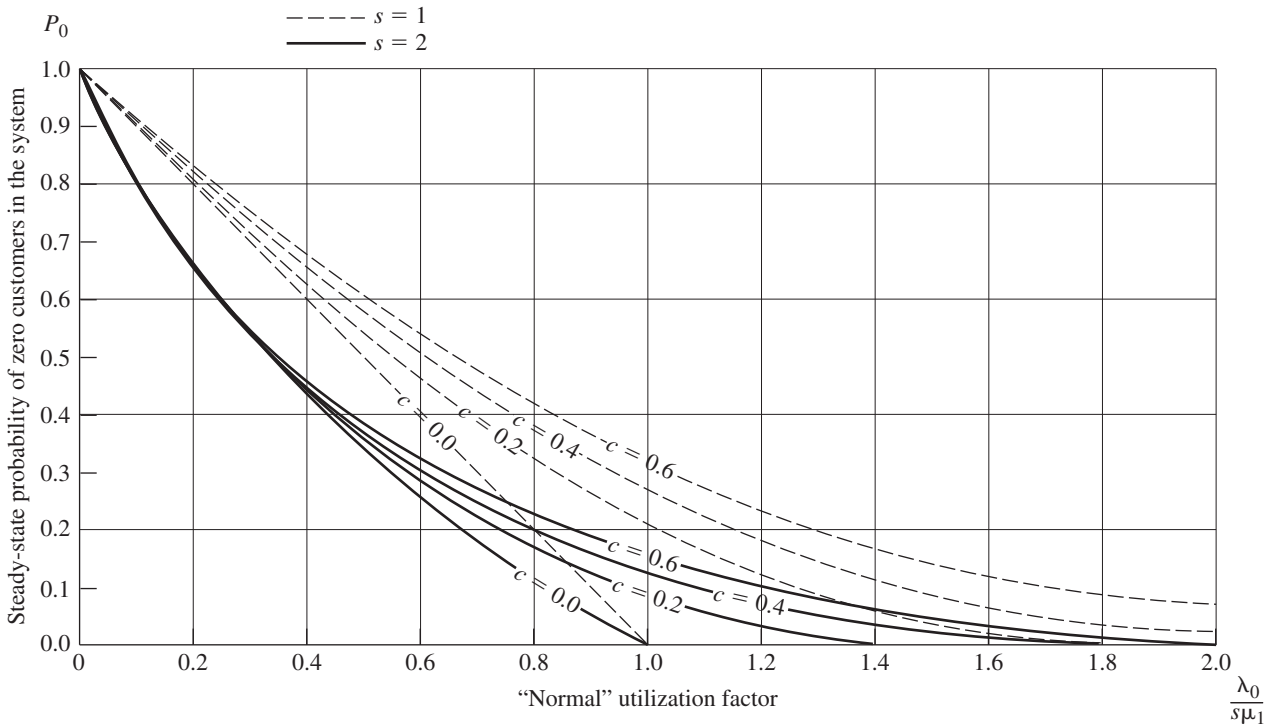


FIGURE 17.9
Values of P_0 for the state-dependent model (Sec. 17.6).

A more general model that combines these two patterns can also be used when both the mean arrival rates and the mean service rates are state-dependent. Thus, let

$$\mu_n = n^a \mu_1 \qquad \text{for } n = 1, 2, \ldots$$

and

$$\lambda_n = (n + 1)^{-b} \lambda_0 \qquad \text{for } n = 0, 1, 2, \ldots .$$

Once again, the C_n values for the birth-and-death process with these parameters are identical to those shown for the state-dependent service rate model when $c = a + b$ and $\lambda/\mu_1 = \lambda_0/\mu_1$, so the tabulated steady-state results actually are applicable to this general model.

Formulation for the Multiple-Server Case ($s > 1$). To generalize this combined model further to the multiple-server case, it seems natural to have the μ_n and λ_n vary with the number of customers *per server* (n/s) in essentially the same way that they vary with n for the single-server case. Thus, let

$$\mu_n = \begin{cases} n\mu_1 & \text{if } n \leq s \\ \left(\dfrac{n}{s}\right)^a s\mu_1 & \text{if } n \geq s \end{cases}$$

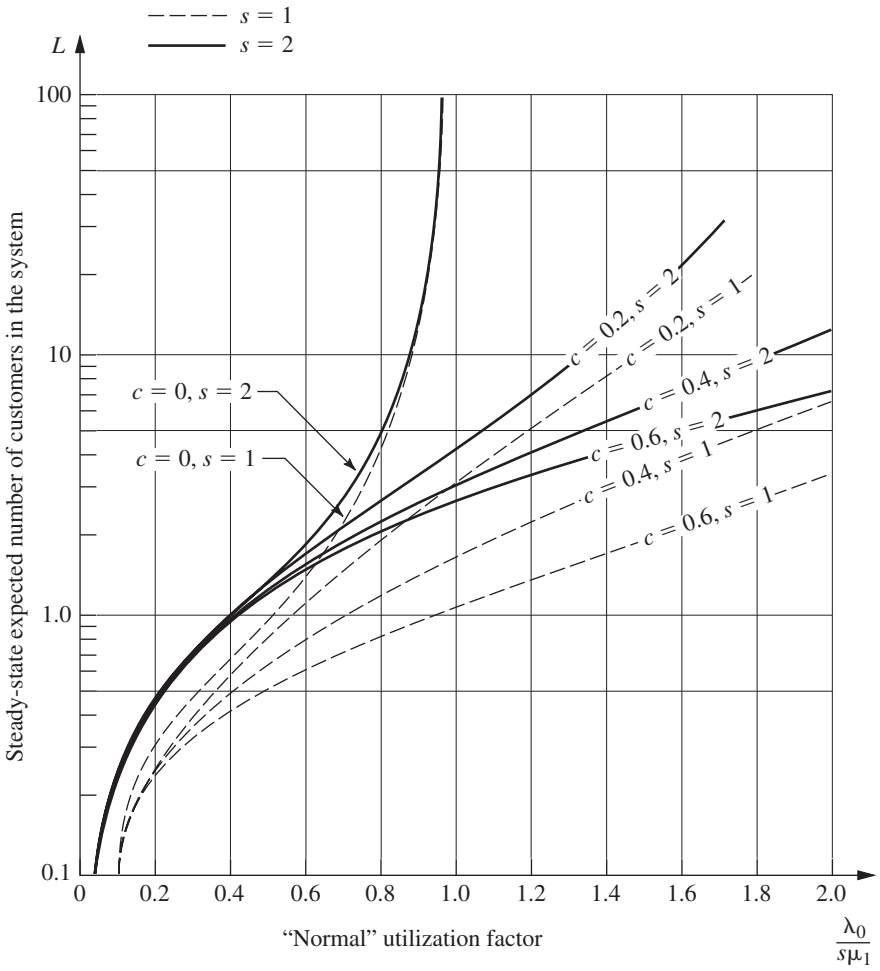


FIGURE 17.10
Values of L for the state-dependent model (Sec. 17.6).

and

$$\lambda_n = \begin{cases} \lambda_0 & \text{if } n \leq s - 1 \\ \left(\dfrac{s}{n+1}\right)^b \lambda_0 & \text{if } n \geq s - 1. \end{cases}$$

Therefore, the birth-and-death process with these parameters has

$$C_n = \begin{cases} \dfrac{(\lambda_0/\mu_1)^n}{n!} & \text{for } n = 0, 1, 2, \ldots, s \\ \dfrac{(\lambda_0/\mu_1)^n}{s!\,(n!/s!)^c\, s^{(1-c)(n-s)}} & \text{for } n = s, s+1, \ldots, \end{cases}$$

where $c = a + b$.

Computational results for P_0, L_q, and L have been tabulated[1] for various values of c, λ_0/μ_1, and s. Some of these results also are given in Figs. 17.9 and 17.10.

The County Hospital Example with State-Dependent Service Rates. After gathering additional data for the County Hospital emergency room, the management engineer found that the time a doctor spends with a patient tends to decrease as the number of patients waiting increases. Part of the explanation is simply that the doctor works faster, but the main reason is that more of the treatment is turned over to a nurse for completion. The pattern of the μ_n (the mean rate at which a doctor treats patients while there are a total of n patients to be treated in the emergency room) seems to fit reasonably the state-dependent service rate model presented here. Therefore, the management engineer has decided to apply this model.

The new data indicate that the average time a doctor spends treating a patient is 24 minutes if no other patients are waiting, whereas this average becomes 12 minutes when each doctor has six patients (so five are waiting their turn). Thus, with a single doctor on duty,

$$\mu_1 = 2\frac{1}{2} \text{ customers per hour,}$$

$$\mu_6 = 5 \text{ customers per hour.}$$

Therefore, the pressure coefficient c (or a in the general model) must satisfy the relationship

$$\mu_6 = 6^c \mu_1, \qquad \text{so} \qquad 6^c = 2.$$

Using logarithms to solve for c yields $c = 0.4$. Because $\lambda = 2$ from before, this solution for c completes the specification of parameter values for this model.

To compare the two alternatives of having one doctor ($s = 1$) or two doctors ($s = 2$) on duty, the management engineer developed the various measures of performance shown in Table 17.3. The values of P_0, L, and (for $s = 2$) L_q were obtained directly from the tabulated results for this model. (Except for this L_q, you can approximate the same values from Figs. 17.9 and 17.10.) These values were then used to calculate

$$\begin{aligned}
P_1 &= C_1 P_0, \\
L_q &= L - (1 - P_0), \qquad \text{if } s = 1, \\
L_q &= L - P_1 - 2(1 - P_0 - P_1), \qquad \text{if } s = 2, \\
W_q &= \frac{L_q}{\lambda}, \qquad W = \frac{L}{\lambda},
\end{aligned}$$

$$P\{W_q > 0\} = 1 - \sum_{n=0}^{s-1} P_n.$$

The fact that some of the results in Table 17.3 do not deviate substantially from those in Table 17.2 reinforces the tentative conclusion that a single doctor will be inadequate next year.

[1]F. S. Hillier, R. W. Conway, and W. L. Maxwell, "A Multiple Server Queueing Model with State Dependent Service Rate," *Journal of Industrial Engineering,* **15:** 153–157, 1964.

TABLE 17.3 Steady-state results from the state-dependent service rate model for the County Hospital problem

	$s = 1$	$s = 2$
$\frac{\lambda}{s\mu_1}$	0.8	0.4
$\frac{\lambda}{s\mu_{6s}}$	0.4	0.2
P_0	0.367	0.440
P_1	0.294	0.352
L_q	0.618	0.095
L	1.251	0.864
W_q	0.309 hour	0.048 hour
W	0.626 hour	0.432 hour
$P\{W_q > 0\}$	0.633	0.208

17.7 QUEUEING MODELS INVOLVING NONEXPONENTIAL DISTRIBUTIONS

Because all the queueing theory models in the preceding section (except for one generalization) are based on the birth-and-death process, both their interarrival and service times are required to have *exponential* distributions. As discussed in Sec. 17.4, this type of probability distribution has many convenient properties for queueing theory, but it provides a reasonable fit for only certain kinds of queueing systems. In particular, the assumption of exponential interarrival times implies that arrivals occur randomly (a Poisson input process), which is a reasonable approximation in many situations but *not* when the arrivals are carefully scheduled or regulated. Furthermore, the actual service-time distribution frequently deviates greatly from the exponential form, particularly when the service requirements of the customers are quite similar. Therefore, it is important to have available other queueing models that use alternative distributions.

Unfortunately, the mathematical analysis of queueing models with nonexponential distributions is much more difficult. However, it has been possible to obtain some useful results for a few such models. This analysis is beyond the level of this book, but in this section we shall summarize the models and describe their results.

The *M*/*G*/1 Model

As introduced in Sec. 17.2, the *M*/*G*/1 model assumes that the queueing system has a *single server* and a *Poisson input process* (exponential interarrival times) with a *fixed* mean arrival rate λ. As usual, it is assumed that the customers have *independent* service times with the *same* probability distribution. However, no restrictions are imposed on what this service-time distribution can be. In fact, it is only necessary to know (or estimate) the mean $1/\mu$ and variance σ^2 of this distribution.

Any such queueing system can eventually reach a steady-state condition if $\rho = \lambda/\mu < 1$. The readily available steady-state results[1] for this general model are the following:

$$P_0 = 1 - \rho,$$
$$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)},$$
$$L = \rho + L_q,$$
$$W_q = \frac{L_q}{\lambda},$$
$$W = W_q + \frac{1}{\mu}.$$

Considering the complexity involved in analyzing a model that permits *any* service-time distribution, it is remarkable that such a simple formula can be obtained for L_q. This formula is one of the most important results in queueing theory because of its ease of use and the prevalence of $M/G/1$ queueing systems in practice. This equation for L_q (or its counterpart for W_q) commonly is referred to as the **Pollaczek-Khintchine formula,** named after two pioneers in the development of queueing theory who derived the formula independently in the early 1930s.

For any fixed expected service time $1/\mu$, notice that L_q, L, W_q, and W all increase as σ^2 is increased. This result is important because it indicates that the consistency of the server has a major bearing on the performance of the service facility—not just the server's average speed. This key point is illustrated in the next subsection.

When the service-time distribution is exponential, $\sigma^2 = 1/\mu^2$, and the preceding results will reduce to the corresponding results for the $M/M/1$ model given at the beginning of Sec. 17.6.

The complete flexibility in the service-time distribution provided by this model is extremely useful, so it is unfortunate that efforts to derive similar results for the multiple-server case have been unsuccessful. However, some multiple-server results have been obtained for the important special cases described by the following two models. (Excel templates are available in this chapter's Excel file for performing the calculations for both the $M/G/1$ model and the two models considered below when $s = 1$.)

The *M/D/s* Model

When the service consists of essentially the same routine task to be performed for all customers, there tends to be little variation in the service time required. The $M/D/s$ model often provides a reasonable representation for this kind of situation, because it assumes that all service times actually equal some fixed *constant* (the *degenerate* service-time distribution) and that we have a *Poisson* input process with a fixed mean arrival rate λ.

[1]A recursion formula also is available for calculating the probability distribution of the number of customers in the system; see A. Hordijk and H. C. Tijms, "A Simple Proof of the Equivalence of the Limiting Distribution of the Continuous-Time and the Embedded Process of the Queue Size in the $M/G/1$ Queue," *Statistica Neerlandica,* **36:** 97–100, 1976.

When there is just a single server, the $M/D/1$ model is just the special case of the $M/G/1$ model where $\sigma^2 = 0$, so that the *Pollaczek-Khintchine formula* reduces to

$$L_q = \frac{\rho^2}{2(1-\rho)},$$

where L, W_q, and W are obtained from L_q as just shown. Notice that these L_q and W_q are exactly *half* as large as those for the exponential service-time case of Sec. 17.6 (the $M/M/1$ model), where $\sigma^2 = 1/\mu^2$, so decreasing σ^2 can *greatly* improve the measures of performance of a queueing system.

For the multiple-server version of this model ($M/D/s$), a complicated method is available[1] for deriving the steady-state probability distribution of the number of customers in the system and its mean [assuming $\rho = \lambda/(s\mu) < 1$]. These results have been tabulated for numerous cases,[2] and the means (L) also are given graphically in Fig. 17.11.

The $M/E_k/s$ Model

The $M/D/s$ model assumes *zero* variation in the service times ($\sigma = 0$), whereas the *exponential* service-time distribution assumes a very large variation ($\sigma = 1/\mu$). Between these two rather extreme cases lies a long middle ground ($0 < \sigma < 1/\mu$), where most *actual* service-time distributions fall. Another kind of theoretical service-time distribution that fills this middle ground is the **Erlang distribution** (named after the founder of queueing theory).

The probability density function for the Erlang distribution is

$$f(t) = \frac{(\mu k)^k}{(k-1)!}\, t^{k-1} e^{-k\mu t}, \qquad \text{for } t \geq 0,$$

where μ and k are strictly positive parameters of the distribution and k is further restricted to be integer. (Except for this integer restriction and the definition of the parameters, this distribution is *identical* to the *gamma distribution*.) Its mean and standard deviation are

$$\text{Mean} = \frac{1}{\mu}$$

and

$$\text{Standard deviation} = \frac{1}{\sqrt{k}}\,\frac{1}{\mu}.$$

Thus, k is the parameter that specifies the degree of variability of the service times relative to the mean. It usually is referred to as the *shape parameter.*

The Erlang distribution is a very important distribution in queueing theory for two reasons. To describe the first one, suppose that $T_1, T_2, \ldots, T_k$ are k independent random variables with an identical exponential distribution whose mean is $1/(k\mu)$. Then their sum

$$T = T_1 + T_2 + \cdots + T_k$$

[1]See N. U. Prabhu: *Queues and Inventories,* Wiley, New York, 1965, pp. 32–34; also see pp. 286–288 in Selected Reference 3.

[2]F. S. Hillier and O. S. Yu, with D. Avis, L. Fossett, F. Lo, and M. Reiman, *Queueing Tables and Graphs,* Elsevier North-Holland, New York, 1981.

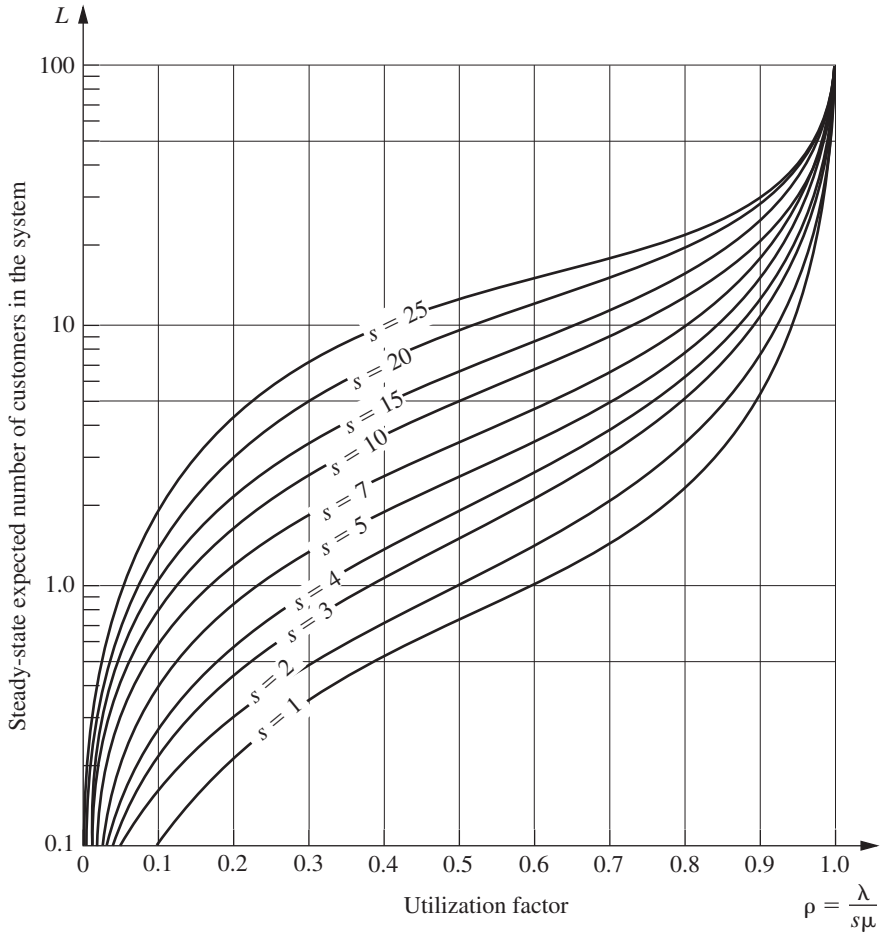


FIGURE 17.11
Values of L for the $M/D/s$ model (Sec. 17.7).

has an *Erlang* distribution with parameters μ and k. The discussion of the exponential distribution in Sec. 17.4 suggested that the time required to perform certain kinds of tasks might well have an exponential distribution. However, the total service required by a customer may involve the server's performing not just one specific task but a sequence of k tasks. If the respective tasks have an identical exponential distribution for their duration, the total service time will have an Erlang distribution. This will be the case, e.g., if the server must perform the *same* exponential task k times for each customer.

The Erlang distribution also is very useful because it is a large (two-parameter) family of distributions permitting only nonnegative values. Hence, empirical service-time distributions can usually be reasonably approximated by an Erlang distribution. In fact, both the *exponential* and the *degenerate* (constant) distributions are special cases of the Erlang distribution, with $k = 1$ and $k = \infty$, respectively. Intermediate values of k provide intermediate distributions with mean $= 1/\mu$, mode $= (k - 1)/(k\mu)$, and variance $= 1/(k\mu^2)$, as

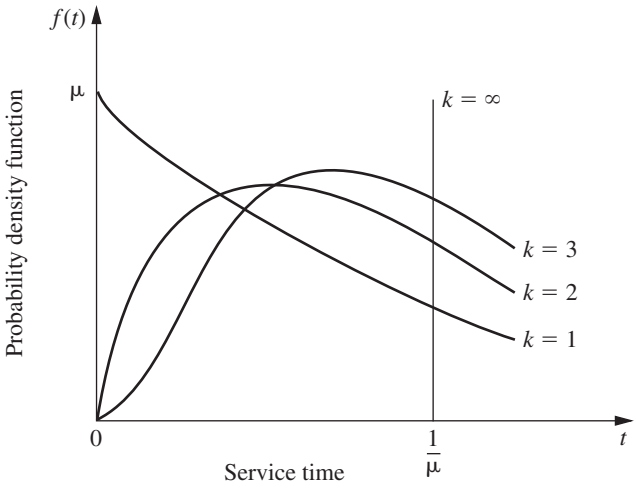


FIGURE 17.12
A family of Erlang distributions with constant mean $1/\mu$.

suggested by Fig. 17.12. Therefore, after estimating the mean and variance of an empirical service-time distribution, these formulas for the mean and variance can be used to choose the integer value of k that matches the estimates most closely.

Now consider the $M/E_k/1$ model, which is just the special case of the $M/G/1$ model where service times have an Erlang distribution with shape parameter $= k$. Applying the Pollaczek-Khintchine formula with $\sigma^2 = 1/(k\mu^2)$ (and the accompanying results given for $M/G/1$) yields

$$L_q = \frac{\lambda^2/(k\mu^2) + \rho^2}{2(1-\rho)} = \frac{1+k}{2k}\frac{\lambda^2}{\mu(\mu-\lambda)},$$

$$W_q = \frac{1+k}{2k}\frac{\lambda}{\mu(\mu-\lambda)},$$

$$W = W_q + \frac{1}{\mu},$$

$$L = \lambda W.$$

With multiple servers ($M/E_k/s$), the relationship of the Erlang distribution to the exponential distribution just described can be exploited to formulate a *modified* birth-and-death process (continuous time Markov chain) in terms of individual exponential service phases (k per customer) rather than complete customers. However, it has not been possible to derive a general steady-state solution [when $\rho = \lambda/(s\mu) < 1$] for the probability distribution of the number of customers in the system as we did in Sec. 17.5. Instead, advanced theory is required to solve individual cases numerically. Once again, these results have been obtained and tabulated for numerous cases.[1] The means (L) also are given graphically in Fig. 17.13 for some cases where $s = 2$.

[1]Ibid.

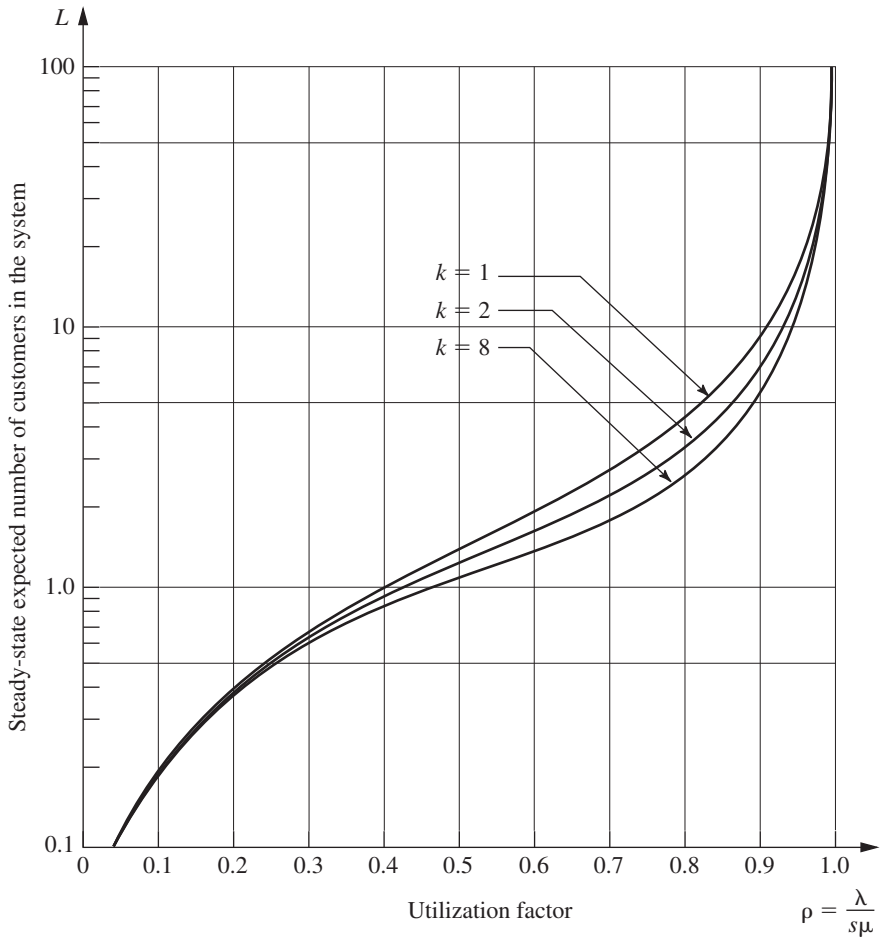


FIGURE 17.13
Values of L for the $M/E_k/2$ model (Sec. 17.7).

Models without a Poisson Input

All the queueing models presented thus far have assumed a Poisson input process (exponential interarrival times). However, this assumption is violated if the arrivals are scheduled or regulated in some way that prevents them from occurring randomly, in which case another model is needed.

As long as the service times have an exponential distribution with a fixed parameter, three such models are readily available. These models are obtained by merely *reversing* the assumed distributions of the *interarrival* and *service times* in the preceding three models. Thus, the first new model ($GI/M/s$) imposes no restriction on what the *interarrival time* distribution can be. In this case, there are some steady-state results available[1] (particularly in regard to waiting-time distributions) for both the single-server and multiple-

[1]For example, see pp. 248–260 of Selected Reference 3.

server versions of the model, but these results are not nearly as convenient as the simple expressions given for the $M/G/1$ model. The second new model ($D/M/s$) assumes that all interarrival times equal some fixed *constant,* which would represent a queueing system where arrivals are *scheduled* at regular intervals. The third new model ($E_k/M/s$) assumes an *Erlang* interarrival time distribution, which provides a middle ground between *regularly scheduled* (constant) and *completely random* (exponential) arrivals. Extensive computational results have been tabulated[1] for these latter two models, including the values of L given graphically in Figs. 17.14 and 17.15.

If neither the interarrival times nor the service times for a queueing system have an exponential distribution, then there are three additional queueing models for which com-

[1]Hillier and Yu, op. cit.

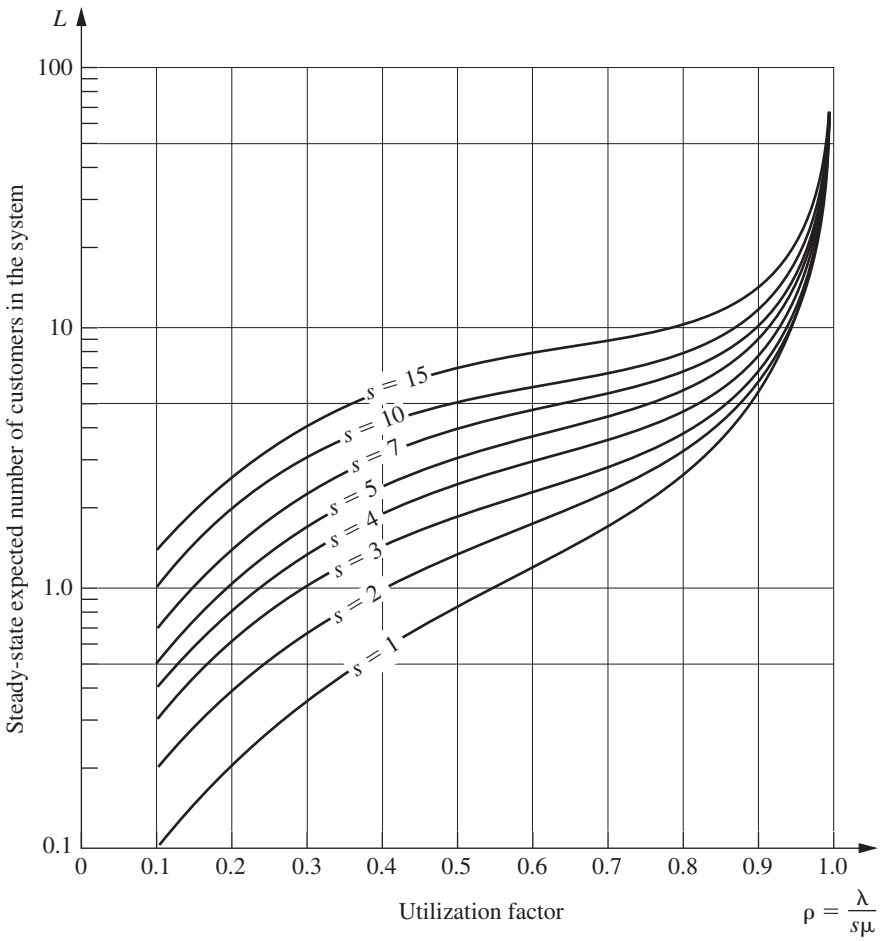


FIGURE 17.14
Values of L for the $D/M/s$ model (Sec. 17.7).

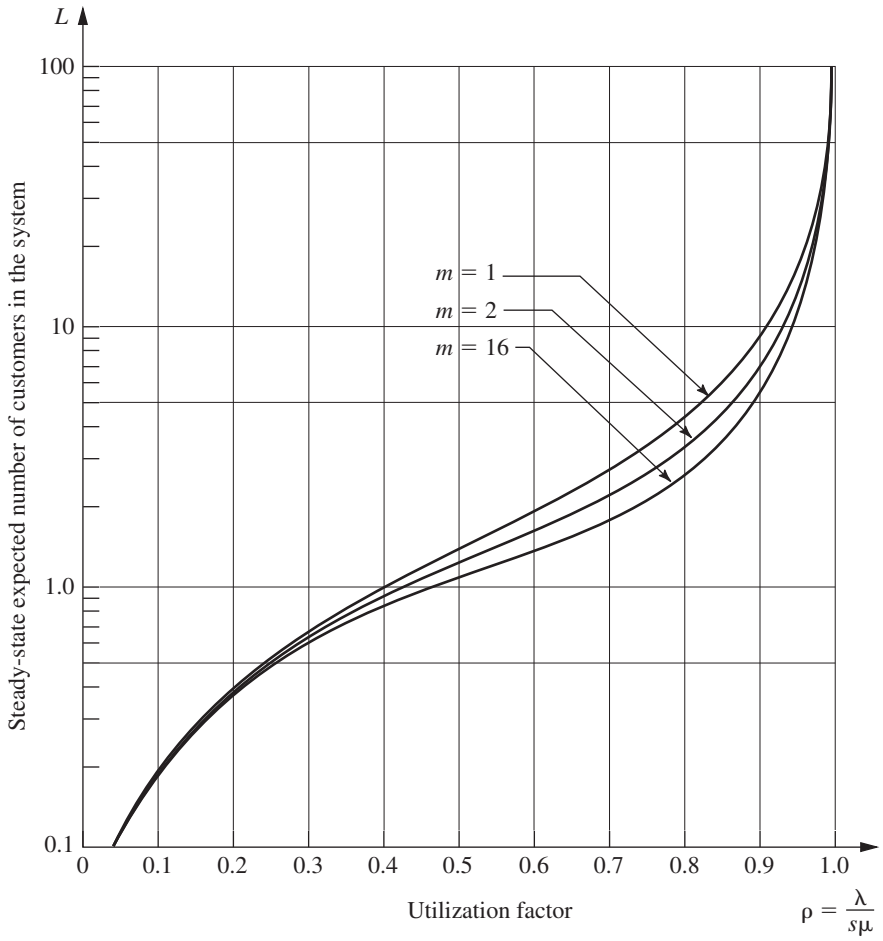


FIGURE 17.15
Values of L for the $E_k/M/2$ model (Sec. 17.7).

putational results also are available.[1] One of these models ($E_m/E_k/s$) assumes an Erlang distribution for both these times. The other two models ($E_k/D/s$ and $D/E_k/s$) assume that one of these times has an Erlang distribution and the other time equals some fixed constant.

Other Models

Although you have seen in this section a large number of queueing models that involve nonexponential distributions, we have far from exhausted the list. For example, another distribution that occasionally is used for either interarrival times or service times is the **hyperexponential distribution.** The key characteristic of this distribution is that even though only nonnegative values are allowed, its standard deviation σ actually is larger than its mean $1/\mu$. This characteristic is in contrast to the Erlang distribution, where $\sigma < 1/\mu$ in every

[1]Ibid.

case except $k = 1$ (exponential distribution), which has $\sigma = 1/\mu$. To illustrate a typical situation where $\sigma > 1/\mu$ can occur, we suppose that the service involved in the queueing system is the repair of some kind of machine or vehicle. If many of the repairs turn out to be routine (small service times) but occasional repairs require an extensive overhaul (very large service times), then the standard deviation of service times will tend to be quite large relative to the mean, in which case the hyperexponential distribution may be used to represent the service-time distribution. Specifically, this distribution would assume that there are fixed probabilities, p and $(1 - p)$, for which kind of repair will occur, that the time required for each kind has an exponential distribution, but that the parameters for these two exponential distributions are different. (In general, the hyperexponential distribution is such a composite of two or more exponential distributions.)

Another family of distributions coming into general use consists of **phase-type distributions** (some of which also are called *generalized Erlangian distributions*). These distributions are obtained by breaking down the total time into a number of phases, each having an exponential distribution, where the parameters of these exponential distributions may be different and the phases may be either in series or in parallel (or both). A group of phases being *in parallel* means that the process randomly selects *one* of the phases to go through each time according to specified probabilities. This approach is, in fact, how the hyperexponential distribution is derived, so this distribution is a special case of the phase-type distributions. Another special case is the Erlang distribution, which has the restrictions that all its k phases are in series and that these phases have the *same* parameter for their exponential distributions. Removing these restrictions means that phase-type distributions in general can provide considerably more flexibility than the Erlang distribution in fitting the actual distribution of interarrival times or service times observed in a real queueing system. This flexibility is especially valuable when using the actual distribution directly in the model is not analytically tractable, and the ratio of the *mean* to the *standard deviation* for the actual distribution does not closely match the available ratios ($\sqrt{k}$ for $k = 1, 2, \ldots$) for the Erlang distribution.

Since they are built up from combinations of exponential distributions, queueing models using phase-type distributions still can be represented by a *continuous time Markov chain.* This Markov chain generally will have an infinite number of states, so solving for the steady-state distribution of the state of the system requires solving an infinite system of linear equations that has a relatively complicated structure. Solving such a system is far from a routine thing, but recent theoretical advances have enabled us to solve these queueing models numerically in some cases. An extensive tabulation of these results for models with various phase-type distributions (including the hyperexponential distribution) is available.[1]

17.8 PRIORITY-DISCIPLINE QUEUEING MODELS

In priority-discipline queueing models, the queue discipline is based on a *priority system.* Thus, the order in which members of the queue are selected for service is based on their assigned priorities.

[1]L. P. Seelen, H. C. Tijms, and M. H. Van Hoorn, *Tables for Multi-Server Queues,* North-Holland, Amsterdam, 1985.

Many real queueing systems fit these priority-discipline models much more closely than other available models. Rush jobs are taken ahead of other jobs, and important customers may be given precedence over others. Therefore, the use of priority-discipline models often provides a very welcome refinement over the more usual queueing models.

We present two basic priority-discipline models here. Since both models make the same assumptions, except for the nature of the priorities, we first describe the models together and then summarize their results separately.

The Models

Both models assume that there are *N priority classes* (class 1 has the highest priority and class *N* has the lowest) and that whenever a server becomes free to begin serving a new customer from the queue, the one customer selected is that member of the *highest* priority class represented in the queue who has waited longest. In other words, customers are selected to begin service in the order of their priority classes, but on a first-come-first-served basis within each priority class. A *Poisson* input process and *exponential* service times are assumed for each priority class. Except for one special case considered later, the models also make the somewhat restrictive assumption that the expected service time is the *same* for all priority classes. However, the models do permit the mean arrival rate to differ among priority classes.

The distinction between the two models is whether the priorities are *nonpreemptive* or *preemptive.* With **nonpreemptive priorities,** a customer being served cannot be ejected back into the queue (preempted) if a higher-priority customer enters the queueing system. Therefore, once a server has begun serving a customer, the service must be completed without interruption. The first model assumes nonpreemptive priorities.

With **preemptive priorities,** the lowest-priority customer being served is *preempted* (ejected back into the queue) whenever a higher-priority customer enters the queueing system. A server is thereby freed to begin serving the new arrival immediately. (When a server does succeed in *finishing* a service, the next customer to begin receiving service is selected just as described at the beginning of this subsection, so a preempted customer normally will get back into service again and, after enough tries, will eventually finish.) Because of the lack-of-memory property of the exponential distribution (see Sec. 17.4), we do not need to worry about defining the point at which service begins when a preempted customer returns to service; the distribution of the *remaining* service time *always* is the same. (For any other service-time distribution, it becomes important to distinguish between *preemptive-resume* systems, where service for a preempted customer resumes at the point of interruption, and *preemptive-repeat* systems, where service must start at the beginning again.) The second model assumes preemptive priorities.

For both models, if the distinction between customers in different priority classes were ignored, Property 6 for the exponential distribution (see Sec. 17.4) implies that *all* customers would arrive according to a Poisson input process. Furthermore, all customers have the *same* exponential distribution for service times. Consequently, the two models actually are identical to the *M*/*M*/*s* model studied in Sec. 17.6 *except* for the order in which customers are served. Therefore, when we count just the *total* number of customers in the system, the steady-state distribution for the *M*/*M*/*s* model also applies to both models. Consequently, the formulas for L and L_q also carry over, as do the expected waiting-time

results (by Little's formula) W and W_q, for a randomly selected customer. What changes is the *distribution* of waiting times, which was derived in Sec. 17.6 under the assumption of a first-come-first-served queue discipline. With a priority discipline, this distribution has a much larger *variance,* because the waiting times of customers in the highest priority classes tend to be much smaller than those under a first-come-first-served discipline, whereas the waiting times in the lowest priority classes tend to be much larger. By the same token, the breakdown of the total number of customers in the system tends to be disproportionately weighted toward the lower-priority classes. But this condition is just the reason for imposing priorities on the queueing system in the first place. We want to *improve* the *measures of performance* for each of the higher-priority classes at the expense of performance for the lower-priority classes. To determine how much improvement is being made, we need to obtain such measures as *expected waiting time in the system* and *expected number of customers in the system* for the individual priority classes. Expressions for these measures are given next for the two models in turn.

Results for the Nonpreemptive Priorities Model

Let W_k be the steady-state expected waiting time in the system (including service time) for a member of priority class k. Then

$$W_k = \frac{1}{AB_{k-1}B_k} + \frac{1}{\mu}, \qquad \text{for } k = 1, 2, \ldots, N,$$

where

$$\begin{aligned}
A &= s!\frac{s\mu - \lambda}{r^s}\sum_{j=0}^{s-1}\frac{r^j}{j!} + s\mu,\\
B_0 &= 1,\\
B_k &= 1 - \frac{\sum_{i=1}^{k}\lambda_i}{s\mu},\\
s &= \text{number of servers},\\
\mu &= \text{mean service rate per busy server},\\
\lambda_i &= \text{mean arrival rate for priority class } i,\\
\lambda &= \sum_{i=1}^{N}\lambda_i,\\
r &= \frac{\lambda}{\mu}.
\end{aligned}$$

(This result assumes that

$$\sum_{i=1}^{k}\lambda_i < s\mu,$$

so that priority class k can reach a steady-state condition.) *Little's formula* still applies to individual priority classes, so L_k, the steady-state expected number of members of priority class k in the queueing system (including those being served), is

$$L_k = \lambda_k W_k, \qquad \text{for } k = 1, 2, \ldots, N.$$

To determine the expected waiting time in the queue (excluding service time) for priority class k, merely subtract $1/\mu$ from W_k; the corresponding expected queue length is again obtained by multiplying by λ_k. For the special case where $s = 1$, the expression for A reduces to $A = \mu^2/\lambda$.

An Excel template is provided in your OR Courseware for performing the above calculations.

A Single-Server Variation of the Nonpreemptive Priorities Model

The above assumption that the expected service time $1/\mu$ is the same for all priority classes is a fairly restrictive one. In practice, this assumption sometimes is violated because of differences in the service requirements for the different priority classes.

Fortunately, for the special case of a single server, it is possible to allow different expected service times and still obtain useful results. Let $1/\mu_k$ denote the mean of the exponential service-time distribution for priority class k, so

$$\mu_k = \text{mean service rate for priority class } k, \qquad \text{for } k = 1, 2, \ldots, N.$$

Then the steady-state expected waiting time in the system for a member of priority class k is

$$W_k = \frac{a_k}{b_{k-1}b_k} + \frac{1}{\mu_k}, \qquad \text{for } k = 1, 2, \ldots, N,$$

where
$$a_k = \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i^2},$$
$$b_0 = 1,$$
$$b_k = 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i}.$$

This result holds as long as

$$\sum_{i=1}^{k} \frac{\lambda_i}{\mu_i} < 1,$$

which enables priority class k to reach a steady-state condition. Little's formula can be used as described above to obtain the other main measures of performance for each priority class.

Results for the Preemptive Priorities Model

For the preemptive priorities model, we need to reinstate the assumption that the expected service time is the same for all priority classes. Using the same notation as for the original nonpreemptive priorities model, having the preemption changes the *total* expected waiting time in the system (including the total service time) to

$$W_k = \frac{1/\mu}{B_{k-1}B_k}, \qquad \text{for } k = 1, 2, \ldots, N,$$

for the *single-server* case ($s = 1$). When $s > 1$, W_k can be calculated by an iterative procedure that will be illustrated soon by the County Hospital example. The L_k continue to satisfy the relationship

$$L_k = \lambda_k W_k, \qquad \text{for } k = 1, 2, \ldots, N.$$

The corresponding results for the queue (excluding customers in service) also can be obtained from W_k and L_k as just described for the case of nonpreemptive priorities. Because of the lack-of-memory property of the exponential distribution (see Sec. 17.4), preemptions do not affect the service process (occurrence of service completions) in any way. The expected total service time for any customer still is $1/\mu$.

This chapter's Excel file includes an Excel template for calculating the above measures of performance for the single-server case.

The County Hospital Example with Priorities

For the County Hospital emergency room problem, the management engineer has noticed that the patients are not treated on a first-come-first-served basis. Rather, the admitting nurse seems to divide the patients into roughly three categories: (1) *critical* cases, where prompt treatment is vital for survival; (2) *serious* cases, where early treatment is important to prevent further deterioration; and (3) *stable* cases, where treatment can be delayed without adverse medical consequences. Patients are then treated in this order of priority, where those in the same category are normally taken on a first-come-first-served basis. A doctor will interrupt treatment of a patient if a new case in a higher-priority category arrives. Approximately 10 percent of the patients fall into the first category, 30 percent into the second, and 60 percent into the third. Because the more serious cases will be sent to the hospital for further care after receiving emergency treatment, the average treatment time by a doctor in the emergency room actually does not differ greatly among these categories.

The management engineer has decided to use a priority-discipline queueing model as a reasonable representation of this queueing system, where the three categories of patients constitute the three priority classes in the model. Because treatment is interrupted by the arrival of a higher-priority case, the *preemptive priorities model* is the appropriate one. Given the previously available data ($\mu = 3$ and $\lambda = 2$), the preceding percentages yield $\lambda_1 = 0.2$, $\lambda_2 = 0.6$, and $\lambda_3 = 1.2$. Table 17.4 gives the resulting expected waiting times in the queue (so *excluding* treatment time) for the respective priority classes[1] when there is one ($s = 1$) or two ($s = 2$) doctors on duty. (The corresponding results for the nonpreemptive priorities model also are given in Table 17.4 to show the effect of preempting.)

Deriving the Preemptive Priority Results. These preemptive priority results for $s = 2$ were obtained as follows. Because the waiting times for priority class 1 customers are completely unaffected by the presence of customers in the lower-priority classes, W_1 will be the same for any other values of λ_2 and λ_3, including $\lambda_2 = 0$ and $\lambda_3 = 0$. There-

[1]Note that these expected times can no longer be interpreted as the expected time before treatment begins when $k > 1$, because treatment may be interrupted at least once, causing additional waiting time before service is completed.

TABLE 17.4 Steady-state results from the priority-discipline models for the County Hospital problem

	Preemptive Priorities		Nonpreemptive Priorities	
	$s = 1$	$s = 2$	$s = 1$	$s = 2$
A	—	—	4.5	36
B_1	0.933	—	0.933	0.967
B_2	0.733	—	0.733	0.867
B_3	0.333	—	0.333	0.667
$W_1 - \frac{1}{\mu}$	0.024 hour	0.00037 hour	0.238 hour	0.029 hour
$W_2 - \frac{1}{\mu}$	0.154 hour	0.00793 hour	0.325 hour	0.033 hour
$W_3 - \frac{1}{\mu}$	1.033 hours	0.06542 hour	0.889 hour	0.048 hour

fore, W_1 must equal W for the corresponding *one-class* model (the $M/M/s$ model in Sec. 17.6) with $s = 2$, $\mu = 3$, and $\lambda = \lambda_1 = 0.2$, which yields

$$W_1 = W = 0.33370 \text{ hour}, \qquad \text{for } \lambda = 0.2$$

so

$$W_1 - \frac{1}{\mu} = 0.33370 - 0.33333 = 0.00037 \text{ hour}.$$

Now consider the first two priority classes. Again note that customers in these classes are completely unaffected by lower-priority classes (just priority class 3 in this case), which can therefore be ignored in the analysis. Let $\overline{W}_{1-2}$ be the expected waiting time in the system (so including service time) of a *random arrival* in *either* of these two classes, so the probability is $\lambda_1/(\lambda_1 + \lambda_2) = \frac{1}{4}$ that this arrival is in class 1 and $\lambda_2/(\lambda_1 + \lambda_2) = \frac{3}{4}$ that it is in class 2. Therefore,

$$\overline{W}_{1-2} = \frac{1}{4}W_1 + \frac{3}{4}W_2.$$

Furthermore, because the *expected* waiting time is the same for *any* queue discipline, $\overline{W}_{1-2}$ must also equal W for the $M/M/s$ model in Sec. 17.6, with $s = 2$, $\mu = 3$, and $\lambda = \lambda_1 + \lambda_2 = 0.8$, which yields

$$\overline{W}_{1-2} = W = 0.33937 \text{ hour}, \qquad \text{for } \lambda = 0.8.$$

Combining these facts gives

$$W_2 = \frac{4}{3}\left[0.33937 - \frac{1}{4}(0.33370)\right] = 0.34126 \text{ hour}.$$

$$\left(W_2 - \frac{1}{\mu} = 0.00793 \text{ hour}.\right)$$

Finally, let $\overline{W}_{1-3}$ be the expected waiting time in the system (so including service time) for a *random arrival* in *any* of the three priority classes, so the probabilities are 0.1, 0.3, and 0.6 that it is in classes 1, 2, and 3, respectively. Therefore,

$$\overline{W}_{1-3} = 0.1W_1 + 0.3W_2 + 0.6W_3.$$

Furthermore, $\overline{W}_{1-3}$ must also equal W for the $M/M/s$ model in Sec. 17.6, with $s = 2$, $\mu = 3$, and $\lambda = \lambda_1 + \lambda_2 + \lambda_3 = 2$, so that (from Table 17.2)

$$\overline{W}_{1-3} = W = 0.375 \text{ hour}, \qquad \text{for } \lambda = 2.$$

Consequently,

$$\begin{aligned} W_3 &= \frac{1}{0.6}\,[0.375 - 0.1(0.33370) - 0.3(0.34126)] \\ &= 0.39875 \text{ hour.} \end{aligned}$$

$$\left(W_3 - \frac{1}{\mu} = 0.06542 \text{ hour.}\right)$$

The corresponding W_q results for the $M/M/s$ model in Sec. 17.6 also could have been used in exactly the same way to derive the $W_k - 1/\mu$ quantities directly.

Conclusions. When $s = 1$, the $W_k - 1/\mu$ values in Table 17.4 for the preemptive priorities case indicate that providing just a single doctor would cause critical cases to wait about $1\frac{1}{2}$ minutes (0.024 hour) on the average, serious cases to wait more than 9 minutes, and stable cases to wait more than 1 hour. (Contrast these results with the average wait of $W_q = \frac{2}{3}$ hour for all patients that was obtained in Table 17.2 under the first-come-first-served queue discipline.) However, these values represent *statistical expectations,* so some patients have to wait considerably longer than the average for their priority class. This wait would not be tolerable for the critical and serious cases, where a few minutes can be vital. By contrast, the $s = 2$ results in Table 17.4 (preemptive priorities case) indicate that adding a second doctor would virtually eliminate waiting for all but the stable cases. Therefore, the management engineer recommended that there be two doctors on duty in the emergency room during the early evening hours next year. The board of directors for County Hospital adopted this recommendation and simultaneously raised the charge for using the emergency room!

17.9 QUEUEING NETWORKS

Thus far we have considered only queueing systems that have a *single* service facility with one or more servers. However, queueing systems encountered in OR studies are sometimes actually *queueing networks,* i.e., networks of service facilities where customers must receive service at some of or all these facilities. For example, orders being processed through a job shop must be routed through a sequence of machine groups (service facilities). It is therefore necessary to study the entire network to obtain such information as the expected total waiting time, expected number of customers in the entire system, and so forth.

Because of the importance of queueing networks, research into this area has been very active. However, this is a difficult area, so we limit ourselves to a brief introduction.

One result is of such fundamental importance for queueing networks that this finding and its implications warrant special attention here. This fundamental result is the following *equivalence property* for the *input process* of arriving customers and the *output process* of departing customers for certain queueing systems.

> **Equivalence property:** Assume that a service facility with s servers and an infinite queue has a Poisson input with parameter λ and the same exponential service-time distribution with parameter μ for each server (the $M/M/s$ model), where $s\mu > \lambda$. Then the steady-state *output* of this service facility is also a Poisson process[1] with parameter λ.

Notice that this property makes no assumption about the type of queue discipline used. Whether it is first-come-first-served, random, or even a priority discipline as in Sec. 17.8, the served customers will leave the service facility according to a Poisson process. The crucial implication of this fact for queueing networks is that if these customers must then go to another service facility for further service, this second facility *also* will have a Poisson input. With an exponential service-time distribution, the equivalence property will hold for this facility as well, which can then provide a Poisson input for a third facility, etc. We discuss the consequences for two basic kinds of networks next.

Infinite Queues in Series

Suppose that customers must all receive service at a *series* of m service facilities in a fixed sequence. Assume that each facility has an infinite queue (no limitation on the number of customers allowed in the queue), so that the series of facilities form a system of *infinite queues in series.* Assume further that the customers arrive at the first facility according to a Poisson process with parameter λ and that each facility i ($i = 1, 2, \ldots, m$) has an exponential service-time distribution with parameter μ_i for its s_i servers, where $s_i\mu_i > \lambda$. It then follows from the equivalence property that (under steady-state conditions) each service facility has a Poisson input with parameter λ. Therefore, the elementary $M/M/s$ model of Sec. 17.6 (or its priority-discipline counterparts in Sec. 17.8) can be used to analyze each service facility independently of the others!

Being able to use the $M/M/s$ model to obtain all measures of performance for each facility independently, rather than analyzing interactions between facilities, is a tremendous simplification. For example, the probability of having n customers at a given facility is given by the formula for P_n in Sec. 17.6 for the $M/M/s$ model. The *joint probability* of n_1 customers at facility 1, n_2 customers at facility 2, . . . , then, is the *product* of the individual probabilities obtained in this simple way. In particular, this joint probability can be expressed as

$$P\{(N_1, N_2, \ldots, N_m) = (n_1, n_2, \ldots, n_m)\} = P_{n_1}P_{n_2}\cdots P_{n_m}.$$

(This simple form for the solution is called the **product form solution.**) Similarly, the expected total waiting time and the expected number of customers in the entire system can be obtained by merely summing the corresponding quantities obtained at the respective facilities.

[1]For a proof, see P. J. Burke: "The Output of a Queueing System," *Operations Research,* **4**(6): 699–704, 1956.

Unfortunately, the equivalence property and its implications do not hold for the case of *finite* queues discussed in Sec. 17.6. This case is actually quite important in practice, because there is often a definite limitation on the queue length in front of service facilities in networks. For example, only a small amount of buffer storage space is typically provided in front of each facility (station) in a production-line system. For such systems of finite queues in series, no simple product form solution is available. The facilities must be analyzed jointly instead, and only limited results have been obtained.

Jackson Networks

Systems of infinite queues in series are not the only queueing networks where the *M*/*M*/*s* model can be used to analyze each service facility independently of the others. Another prominent kind of network with this property (a product form solution) is the *Jackson network,* named after the individual who first characterized the network and showed that this property holds.[1]

The characteristics of a Jackson network are the same as assumed above for the system of infinite queues in series, except now the customers visit the facilities in different orders (and may not visit them all). For each facility, its arriving customers come from *both* outside the system (according to a Poisson process) and the other facilities. These characteristics are summarized below.

A **Jackson network** is a system of m service facilities where facility i ($i = 1, 2, \ldots, m$) has

1. An infinite queue
2. Customers arriving from outside the system according to a Poisson input process with parameter a_i
3. s_i servers with an exponential service-time distribution with parameter μ_i.

A customer leaving facility i is routed next to facility j ($j = 1, 2, \ldots, m$) with probability p_{ij} or departs the system with probability

$$q_i = 1 - \sum_{j=1}^{m} p_{ij}.$$

Any such network has the following key property.

Under steady-state conditions, each facility j ($j = 1, 2, \ldots, m$) in a Jackson network behaves as if it were an *independent M*/*M*/*s* queueing system with arrival rate

$$\lambda_j = a_j + \sum_{i=1}^{m} \lambda_i p_{ij},$$

where $s_j\mu_j > \lambda_j$.

This key property cannot be *proved* directly from the equivalence property this time (the reasoning would become circular), but its *intuitive underpinning* is still provided by the latter property. The intuitive viewpoint (not quite technically correct) is that, for each facility i, its input processes from the various sources (outside and other facilities) are *independent Poisson processes,* so the *aggregate* input process is Poisson with parameter λ_i (Prop-

[1]See J. R. Jackson, "Jobshop-Like Queueing Systems," *Management Science,* **10**(1): 131–142, 1963.

erty 6 in Sec. 17.4). The equivalence property then says that the *aggregate output* process for facility i must be Poisson with parameter λ_i. By disaggregating this output process (Property 6 again), the process for customers going from facility i to facility j must be Poisson with parameter $\lambda_i p_{ij}$. This process becomes one of the Poisson *input* processes for facility j, thereby helping to maintain the series of Poisson processes in the overall system.

The equation given for obtaining λ_j is based on the fact that λ_i is the *departure rate* as well as the arrival rate for all customers using facility i. Because p_{ij} is the proportion of customers departing from facility i who go next to facility j, the rate at which customers from facility i arrive at facility j is $\lambda_i p_{ij}$. Summing this product over all i, and then adding this sum to a_j, gives the *total arrival rate* to facility j from all sources.

To calculate λ_j from this equation requires knowing the λ_i for $i \neq j$, but these λ_i also are unknowns given by the corresponding equations. Therefore, the procedure is to solve *simultaneously* for $\lambda_1, \lambda_2, \ldots, \lambda_m$ by obtaining the simultaneous solution of the entire system of linear equations for λ_j for $j = 1, 2, \ldots, m$. Your OR Courseware includes an Excel template for solving for the λ_j in this way.

To illustrate these calculations, consider a Jackson network with three service facilities that have the parameters shown in Table 17.5. Plugging into the formula for λ_j for $j = 1, 2, 3$, we obtain

$$\begin{aligned}
\lambda_1 &= 1 + 0.1\lambda_2 + 0.4\lambda_3 \\
\lambda_2 &= 4 + 0.6\lambda_1 + 0.4\lambda_3 \\
\lambda_3 &= 3 + 0.3\lambda_1 + 0.3\lambda_2.
\end{aligned}$$

(Reason through each equation to see why it gives the total arrival rate to the corresponding facility.) The simultaneous solution for this system is

$$\lambda_1 = 5, \qquad \lambda_2 = 10, \qquad \lambda_3 = 7\frac{1}{2}.$$

Given this simultaneous solution, each of the three service facilities now can be analyzed *independently* by using the formulas for the *M*/*M*/*s* model given in Sec. 17.6. For example, to obtain the distribution of the number of customers $N_i = n_i$ at facility i, note that

$$\rho_i = \frac{\lambda_i}{s_i \mu_i} = \begin{cases} \frac{1}{2} & \text{for } i = 1 \\ \frac{1}{2} & \text{for } i = 2 \\ \frac{3}{4} & \text{for } i = 3. \end{cases}$$

TABLE 17.5 Data for the example of a Jackson network

Facility j	s_j	μ_j	a_j	p_{ij}		
				$i = 1$	$i = 2$	$i = 3$
$j = 1$	1	10	1	0	0.1	0.4
$j = 2$	2	10	4	0.6	0	0.4
$j = 3$	1	10	3	0.3	0.3	0

Plugging these values (and the parameters in Table 17.5) into the formula for P_n gives

$$P_{n_1} = \frac{1}{2}\left(\frac{1}{2}\right)^{n_1} \qquad \text{for facility 1,}$$

$$P_{n_2} = \begin{cases} \dfrac{1}{3} & \text{for } n_2 = 0 \\ \dfrac{1}{3} & \text{for } n_2 = 1 \\ \dfrac{1}{3}\left(\dfrac{1}{2}\right)^{n_2 - 1} & \text{for } n_2 \geq 2 \end{cases} \qquad \text{for facility 2,}$$

$$P_{n_3} = \frac{1}{4}\left(\frac{3}{4}\right)^{n_3} \qquad \text{for facility 3.}$$

The *joint probability* of (n_1, n_2, n_3) then is given simply by the product form solution

$$P\{(N_1, N_2, N_3) = (n_1, n_2, n_3)\} = P_{n_1}P_{n_2}P_{n_3}.$$

In a similar manner, the expected number of customers L_i at facility i can be calculated from Sec. 17.6 as

$$L_1 = 1, \qquad L_2 = \frac{4}{3}, \qquad L_3 = 3.$$

The expected *total* number of customers in the entire system then is

$$L = L_1 + L_2 + L_3 = 5\frac{1}{3}.$$

Obtaining W, the expected *total* waiting time in the system (including service times) for a customer, is a little trickier. You cannot simply add the expected waiting times at the respective facilities, because a customer does not necessarily visit each facility exactly once. However, Little's formula can still be used, where the system arrival rate λ is the sum of the arrival rates *from outside* to the facilities, $\lambda = a_1 + a_2 + a_3 = 8$. Thus,

$$W = \frac{L}{a_1 + a_2 + a_3} = \frac{2}{3}.$$

In conclusion, we should point out that there do exist other (more complicated) kinds of queueing networks where the individual service facilities can be analyzed independently from the others. In fact, finding queueing networks with a product form solution has been the Holy Grail for research on queueing networks. Two sources of additional information are Selected References 6 and 7.

17.10 CONCLUSIONS

Queueing systems are prevalent throughout society. The adequacy of these systems can have an important effect on the quality of life and productivity.

Queueing theory studies queueing systems by formulating mathematical models of their operation and then using these models to derive measures of performance. This analysis provides vital information for effectively designing queueing systems that achieve an appropriate balance between the cost of providing a service and the cost associated with waiting for that service.

This chapter presented the most basic models of queueing theory for which particularly useful results are available. However, many other interesting models could be considered if space permitted. In fact, several thousand research papers formulating and/or analyzing queueing models have already appeared in the technical literature, and many more are being published each year!

The *exponential distribution* plays a fundamental role in queueing theory for representing the distribution of interarrival and service times, because this assumption enables us to represent the queueing system as a *continuous time Markov chain.* For the same reason, *phase-type distributions* such as the *Erlang distribution,* where the total time is broken down into individual phases having an exponential distribution, are very useful. Useful analytical results have been obtained for only a relatively few queueing models making other assumptions.

Priority-discipline queueing models are useful for the common situation where some categories of customers are given priority over others for receiving service.

In another common situation, customers must receive service at several different service facilities. Models for queueing networks are gaining widespread use for such situations. This is an area of especially active ongoing research.

When no tractable model that provides a reasonable representation of the queueing system under study is available, a common approach is to obtain relevant performance data by developing a computer program for simulating the operation of the system. This technique is discussed in Chap. 22.

Chapter 18 describes how queueing theory can be used to help design effective queueing systems.

SELECTED REFERENCES

1. Cooper, R. B.: *Introduction to Queueing Theory,* 2d ed., Elsevier North-Holland, New York, 1981. (Also distributed by the George Washington University Continuing Engineering Education Program, Washington, DC.)
2. Cooper, R. B.: "Queueing Theory," Chap. 10 in D. P. Heyman and M. J. Soble (eds.), *Stochastic Models,* North Holland, Amsterdam and New York, 1990. (This survey paper also is distributed by the George Washington University Continuing Engineering Education Program, Washington, DC.)
3. Gross, D., and C. M. Harris: *Fundamentals of Queueing Theory,* 3d ed., Wiley, New York, 1998.
4. Kleinrock, L.: *Queueing Systems, Vol. I: Theory,* Wiley, New York, 1975.
5. Prabhu, N. U.: *Foundations of Queueing Theory,* Kluwer Academic Publishers, Boston, 1997.
6. van Dijk, N. M.: *Queueing Networks and Product Forms: A Systems Approach,* Wiley, New York, 1993.
7. Walrand, J.: *An Introduction to Queueing Networks,* Prentice-Hall, Englewood Cliffs, NJ, 1988.
8. Wolff, R. W.: *Stochastic Modeling and the Theory of Queues,* Prentice-Hall, Englewood Cliffs, NJ, 1989.

LEARNING AIDS FOR THIS CHAPTER IN YOUR OR COURSEWARE

"Ch. 17—Queueing Theory" Excel File:

Template for *M*/*M*/*s* Model
Template for Finite Queue Variation of *M*/*M*/*s* Model