4

Introduction to Queueing Theory and Its Applications

Requests for service very often find some or all of the units associated with some urban service system busy. The performance of these systems on such occasions (i.e., under conditions of some "stress") plays a major role in shaping citizens' perceptions with regard to the level of service offered.

That some deterioration in the level of service will occur during periods of intensive activity is clear. For instance, a fire alarm may occur at an instant when all fire companies stationed at the nearest fire houses have already been dispatched to another alarm elsewhere in a city. To the extent, then, that the response time to the new fire will be longer under these circumstances (because of the need to dispatch fire engines from remote fire stations), the perceived level of service will be lower. Similarly, a police dispatcher will often postpone service to medium-priority calls for police assistance during periods when nearly all police cars in a district are busy. In doing so, the dispatcher preserves the ability of the police to respond immediately to top-priority calls while maintaining some level of police "visibility" through the patrolling activities of nonbusy cars. Under normal conditions such medium-priority calls would have received prompt service.

Busy period service delays must inevitably occur in the case of services that respond to unpredictable demands whose time and location of occurrence are governed by some type of (known or unknown) probabilistic law. The cost of providing sufficient capacity to avoid all delays under all circumstances would be insupportable. The proper role of analysis, therefore, is to design service systems that achieve an acceptable balance between system operating costs, on the one hand, and the delays suffered by users of that system, on the other. As to where this acceptable balance lies, it all depends

on the nature of the service provided. In some cases—for example, in fire or ambulance services—the costs of delays are generally perceived to be very high. To assure a low probability of such delays, it is therefore necessary to design systems that are relatively underutilized and whose servers (e.g., fire engines, ambulances, etc.), experience long periods of idleness. In other instances (e.g., collection of solid refuse or the delivery of mail), delays of a few hours or even of days are not usually catastrophic. This, in turn, permits a high level of utilization for the servers used by the system.

Queueing theory, the theory of congestion, is the branch of operations research which explores the relationships between demand on a service system and the delays suffered by the users of that system. Since almost all urban service systems can be viewed as queueing systems (as it will become clear in this chapter), queueing theory plays a central role in the analysis of and planning for urban services. This chapter will therefore deal with a review of some important results in queueing theory and with an introduction to the applications of these results to the problems on which this book focuses.

4.1 QUESTIONS AND ANSWERS IN QUEUEING THEORY

It is important that those who wish to apply the results of queueing theory have an appreciation for the kinds of questions that queueing theory can answer and for the nature of and the assumptions behind these answers.

In working with queueing theory one must, first of all, take the particular real-world system of interest, study this system, and create (or simply choose from the list of models in queueing theory) a mathematical model to represent it. Through the analysis of this mathematical model, one then obtains the answers, which supposedly apply to the original system as well. Inherent to the procedure of creating a mathematical model are the notions of *simplification* and *approximation*: The analyst must necessarily disregard many details which he or she considers superfluous (or of minor importance) to the central points of interest. In most cases, approximations must also be made in transforming raw and often incomplete data into mathematical quantities that will make the analysis of the model possible. Finally, it is not unusual for an analyst to make many assumptions about certain aspects of the behavior of the real system—assumptions based mostly on intuition and experience rather than on any real evidence that the system indeed behaves in this way. Under the circumstances it would then be fair to state that, in most applications, the estimates of quantities of interest which can be obtained through a queueing analysis should only be viewed as approximate indicators of the size of these quantities in the real world. Consequently, the application of queueing theory is most useful in pointing out the inadequacies of existing

operating systems, the directions in which to proceed for improving these systems, and the approximate values that some of the controllable variables of the system must assume to achieve a satisfactory level of performance.

A second major point that should be realized is that queueing theory does not offer a full menu of answers. The state-of-the-art after nearly three decades of intensive research can be summarized roughly as follows:

1. Few closed-form expressions exist for the transient and the nonstationary behavior of queueing systems. Almost all the existing important results of queueing theory are obtained for equilibrium conditions (i.e., with the queueing system operating in the "steady state," in engineering parlance).
2. Even assuming equilibrium conditions, queueing theory runs into enormous mathematical difficulties in all but relatively few types of situations. Quite often, the choice facing an analyst is between, on the one hand, using a realistic mathematical model for which almost no results can be obtained and, on the other, using a simplified model that provides results of questionable validity for the problem at hand.
3. Most of the exact results of queueing theory apply to queueing systems in which the interarrival times or the service times or (ideally) both are negative exponential. Fortunately, there are many real-world systems for which at least the interarrival times are negative exponential. The main reason is that many arrival processes observable in practice can be modeled as Poisson processes (which in turn implies negative exponential interarrival times). This is especially true when one refers to *urban* service systems, where Poisson (or nearly Poisson) arrival processes are abundant.
4. Queueing theory is "very good" at estimating the low moments and central moments of such important quantities as the waiting times or the "number of users present" in queueing systems but not nearly as good at computing the probability distributions for these quantities. Indeed, in all but a handful of cases, the only approach available for obtaining probability distributions for most of the quantities of interest in queueing theory is through the use of a combination of transform analysis and numerical analysis techniques. We shall see examples of the numerical analysis approach later in this chapter and in Chapter 5.

Items (1)–(4) above, discouraging as they may sound, are only meant to provide some perspective and not to detract from the value of the results that queueing theory has generated to date. In fact, some of these results are very powerful. They apply to quite general queueing systems and provide

important information about the queueing phenomena that occur, while requiring only a minimum amount of knowledge about the characteristics of interarrival times, service times, queue discipline, and so on.

In the following sections, our attention will be focused on answering questions related to estimating such quantities as the fraction of time service facilities are idle (or busy); the expected values (and occasionally the variance and simple moments) of the time spent by queueing system users while waiting to gain access to servers; the expected duration of periods during which a server is continually busy; and the number of other users that an arriving user can expect to find in a queueing system. As we have just indicated, these are precisely the types of questions that queueing theory has been most successful with. In the process we shall present, whenever available, other results which cast additional light on the queueing phenomena that we wish to explore.

4.2 BACKGROUND, TERMS, AND SOME CONVENTIONAL NOTATION

In this chapter we shall use the term *queueing system* to refer to a generic model (see Figure 4.1) which comprises three elements: a user source, a queue, and a service facility that contains one or more (including possibly an infinite number of) *identical servers in parallel*. Thus, each user of a queueing system is "generated" by a user source, passes through the queue where (s)he may remain for a nonnegative period of time (including possibly zero time), and then is processed by a single server—because of the parallel arrangement of servers. Once a user has left any one of the servers in the system, after obtaining service there, the user is considered to have left the queueing system as well.

In view of the description given above, a *queueing network* can now be defined as a set of interconnected queueing systems. Thus, in a queueing network, the user sources for some of the queueing systems in the network may be other queueing systems in the same network (see also Figure 4.2). It can also be inferred that the analysis of models of queueing systems, as we have defined them here, provides the building blocks for the analysis of queueing networks.

To describe a queueing system fully, information must be supplied about all three generic elements of the system (Figure 4.1):

1. About the user generating process (i.e., the arrival process of users at the system).
2. About the queue discipline [i.e., the order in which users obtain access to the service facility, once (and if) they join the queue].
3. About the service process.

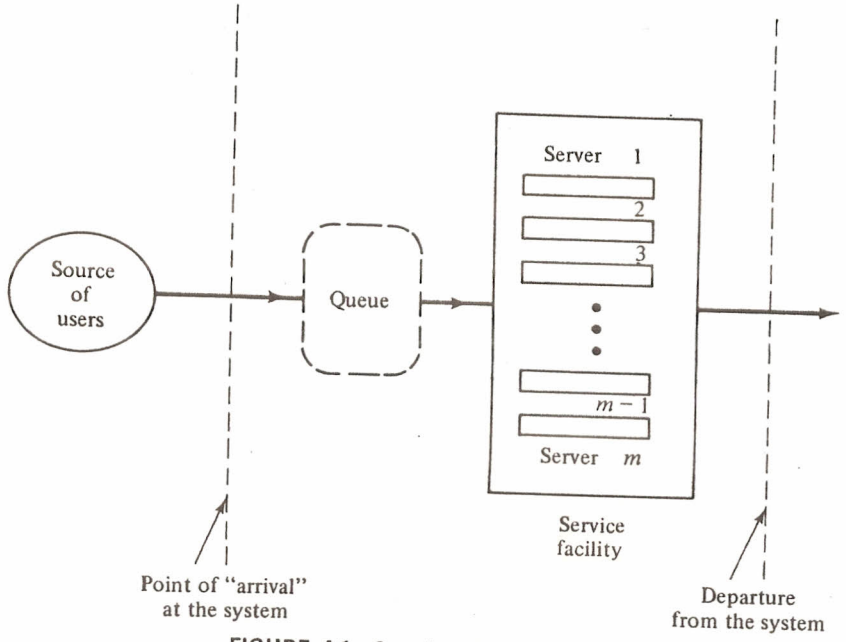

FIGURE 4.1 Generic queueing system.

To describe a queueing *network*, further information must be provided (Figure 4.2) on how the queueing systems are interconnected, how they interact (e.g., do bottlenecks "downstream" ever affect the preceding servers?), and how users are assigned to the queueing systems.

It should be obvious that there exist countless variations of queueing systems and networks. We shall have occasion to refer to some of these variations in following sections. At this point, however, it should be noted that a code has been used in queueing theory for years to describe some of

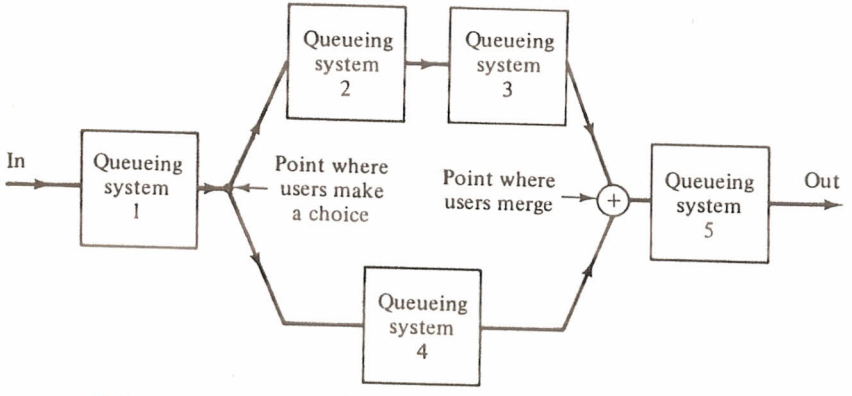

FIGURE 4.2 Queueing network consisting of five queueing systems.

the simplest (and best understood) queueing systems. This code is of the form $A/B/m$, where A and B are letter symbols and m is an integer constant. The letters A and B indicate the probability distribution of user interarrival times and of service times, respectively, and m is the number of identical parallel servers in the queueing system (thus, m can take values from 1 to ∞).

The standard code letters used for probability distributions in queueing theory are:

M = Poisson (i.e., negative exponential pdf for user interarrival times or for service times; M stands for "memoryless")

D = deterministic (i.e., interarrival or service times are constant)

E_k = kth-order Erlang distribution [see (2.50) and (2.57)]

H_k = kth-order hyperexponential distribution (see Problem 4.6)

G = "general" distribution (i.e., any distribution at all)

The letters A and B can thus be any one of the five symbols above.

To newcomers to queueing theory it always seems strange that, out of all possible probability distributions, only four (M, D, E_k, and H_k) have been assigned special symbols. The simple reason is that only these four distributions offer significant advantages in an abstract analysis, in the sense that when they are present in a queueing model, parts of the mathematical analysis of that model may become more tractable. Thus, all other distributions are lumped under the "general" (G) category (which, of course, also includes the M, D, E_k, and H_k cases). These comments will become more clear as we make our way through this chapter.

The coded systems also assume *independence* of successive user arrival times and of successive service times at the queueing system. We shall see several examples of urban service systems where this assumption is not valid in practice.

Some more-or-less standard abbreviations are also used to indicate the most commonly encountered queue disciplines. FIFO is used to indicate the first-in, first-out queueing arrangement, also known as FCFS (= first come, first served). Similarly, LIFO (= last in, first out) or LCFS (= last come, first served) indicate the situation in which the last user to join the queue becomes the next in line for entering service. The abbreviation SIRO is used to indicate "service in random order." Although LIFO and SIRO may at times appear offensive to our sense of fairness—especially in connection with gaining access to public transportation vehicles during peak traffic hours—they, and their variations are all too common in real life and thus cannot be ignored.

Another important parameter in the description of a queueing system is

the *system capacity*. This indicates the maximum number of users that can at any time be in the service facility *and* in the queue. *Queue capacity*, on the other hand, indicates the maximum number of users that can be in the queue alone.

Finally, it is important to emphasize that queueing systems need not adhere to the classical picture of a physically stationary server (a bank teller, a checkout counter, highway toll booths, etc.) where prospective users queue up to be served. It is very common to have systems in which the prospective users remain stationary, possibly at geographically widely separated points, while the server(s) associated with the queueing system visit them according to some (implicit or explicit) priority scheme and provide the requested service. We often use the term *spatially distributed queues* to refer to such geographically "spread-out" systems. They are of particular interest in urban operations research and Chapter 5 will deal with their specific characteristics. For some of these systems (e.g., fire departments or ambulance services), queues in the sense of a backlog of calls for service, rarely, if ever, occur. Yet queueing theory is most valuable in planning for such systems—in determining, for instance, the number of vehicles needed, the expected workload of service units, or the best deployment of the service units in a city.

4.3 DEFINING THE QUANTITIES OF INTEREST

In this section we define the quantities and introduce the notation that will be used in the rest of the chapter. We do this by focusing on a specific queueing system and beginning to count at some instant $t = 0$ the number of users who arrive at the system.

Let us concentrate on the ith user to arrive at that system after we begin our counting process. Three important "events" can be identified with respect to this ith user: the arrival of the user at the queueing system, the beginning of service to the user, and the completion of service to the user. We shall denote the instants when these three events occur as $t_a(i)$, $t_b(i)$, and $t_c(i)$, respectively (with a standing for "arrival," b for "beginning" of service, and c for "completion" of service).

We can now define the following quantities:

$x(i) \triangleq t_a(i) - t_a(i-1)$ = ith interarrival time

$s(i) \triangleq t_c(i) - t_b(i)$ = service time for the ith (in terms of order of arrival to the system) user

$w_q(i) \triangleq t_b(i) - t_a(i)$ = waiting time in the queue for the ith user

$w(i) \triangleq t_c(i) - t_a(i)$ = total time spent in the queueing system by the ith user ("system occupancy time")

Obviously, from the foregoing definitions we also have

$$w(i) = w_q(i) + s(i) \tag{4.1}$$

In general, the interarrival and service times are random variables, $X(i)$ and $S(i)$, with pdf's $f_{X(i)}(x)$ and $f_{S(i)}(s)$, respectively.[1] We shall assume from now on that, unless otherwise stated, the interarrival-time random variables $X(i)$ are independent and identically distributed [i.e., $f_{X(1)}(x) = f_{X(2)}(x) = \ldots = f_X(x)$]. Similarly, we shall assume that the service times $S(i)$ are independent and identically distributed with $f_{S(1)}(s) = f_{S(2)}(s) = \ldots = f_S(s)$. The expected values of random variables X and S appear so frequently in the analysis of queueing systems that special symbols have been adopted for them:

$$\frac{1}{\lambda} \triangleq E[X] \tag{4.2}$$

$$\frac{1}{\mu} \triangleq E[S] \tag{4.3}$$

In words, λ represents the expected number (or the "rate") of user arrivals at the queueing system per unit of time. Similarly, μ is the expected number of service completions per unit of time when a server is working continuously. Note that when all m parallel and identical servers in a queueing system are working simultaneously, the rate of service completions is equal to $m\mu$.

Now if the queueing system is allowed to operate for a long time, it can be expected, under certain conditions, to reach an equilibrium ("steady state"). Without specifying what these conditions are, it is reasonable to assume that the system occupancy times, $w(i)$, and waiting times, $w_q(i)$, for large values of i, will tend to become samples of two random variables W and W_q, respectively, whose pdf's $f_W(w)$ and $f_{W_q}(w_q)$ are independent of the order, i, of a user's arrival. We shall refer to $f_W(w)$ and $f_{W_q}(w_q)$ as the steady-state probability density functions for the system occupancy times and the waiting times of users, respectively. We shall then define the quantities

$\bar{W} \triangleq E[W] = \lim_{i\to\infty} E[W(i)]$ = expected system occupancy time for a user under steady-state conditions

$\bar{W}_q \triangleq E[W_q] = \lim_{i\to\infty} E[W_q(i)]$ = expected waiting time in queue for a user under steady-state conditions

Rather than focus on user-related events at the queueing system [such as the $t_a(i)$, $t_b(i)$, $t_c(i)$, etc.], we also could have looked at the system at random

[1] The use of the term "probability density function" is made here in a general sense. The random variables $X(i)$ and $S(i)$, $i = 1, 2, 3, \ldots$ can be discrete, continuous, or mixed.

points in time and defined the quantities

$N(t) \triangleq$ total number of users (including those in service) who are in the queueing system at time t

$N_q(t) \triangleq$ number of users waiting in the queue at time t

For large values of t and under the (yet unspecified) proper conditions, we can expect the distributions of variables $N(t)$ and $N_q(t)$ to approach equilibrium (steady-state) pmf's $p_N(n)$ and $p_{N_q}(n)$.

We shall then use the symbols[2]

$$\bar{L} \triangleq E[N] = \lim_{t\to\infty} E[N(t)] = \text{expected total number of users in the queueing system under steady-state conditions}$$

$$\bar{L}_q \triangleq E[N_q] = \lim_{t\to\infty} E[N_q(t)] = \text{expected number of users in the queue under steady-state conditions}$$

We shall also define here the quantity

$$\rho \triangleq \text{"utilization ratio"} = \frac{\text{rate of user arrivals at a queueing system}}{\text{total available rate of service by the service facility}}$$

From the definition of the utilization ratio (the reason for its name will become obvious later), it is clear that, for a single-server queueing system,

$$\rho = \frac{\lambda}{\mu} = \lambda \cdot E[S] \tag{4.4}$$

whereas for a m-server queueing system,

$$\rho = \frac{\lambda}{m\mu} = \frac{\lambda E[S]}{m} \tag{4.5}$$

4.4 SOME IMPORTANT RELATIONSHIPS IN QUEUEING THEORY

We mentioned in Section 4.1 that queueing theory has been highly successful in deriving expressions for the low moments of quantities such as the waiting times of the users of a queueing system or the number of users present in the system at a given time. We shall now begin our discussion with the derivation of a few important relationships involving the expected values of these

[2]The notation $\bar{L}, \bar{L}_q, \bar{W}$, and $\bar{W}_q$ to denote expected values is unusual for this book. We employ it to be consistent with notation that has become more-or-less standard in queueing theory.

quantities, $\bar{L}, \bar{L}_q, \bar{W}$, and $\bar{W}_q$, for a very general queueing system. J. D. C. Little [LITT 61] is generally credited with being the first to prove these relationships formally. Subsequent authors have shown that Little's results are valid for queueing systems more general than he assumed in his original work [STID 74]. Here we shall use an informal and intuitive argument paralleling the one offered by Kleinrock [KLEI 75].

The search for the relationships is motivated by the intuition-satisfying notion that the average length of the queue in front of a facility offers a good indication of the average waiting time for use of that facility (and vice versa). These relationships turn out to be especially simple.

First, we shall position ourselves at the "entrance" of a queueing system and count the number of users that arrive there during an interval of arbitrary length, τ, beginning at the time $t = 0$ when the system is empty (see also Figure 4.3). We let

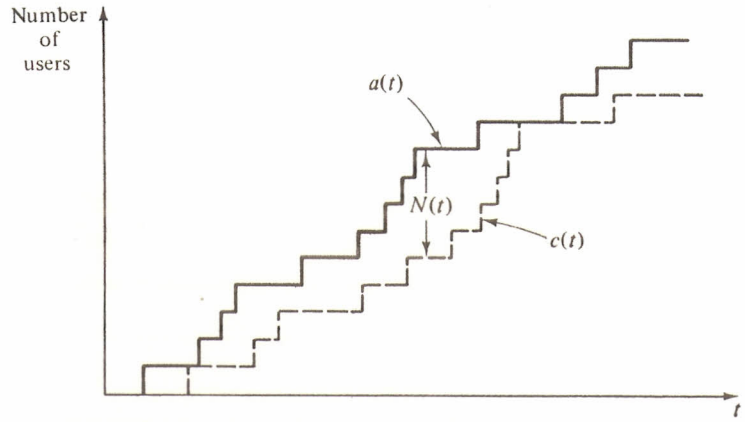

FIGURE 4.3 Arrivals and service completions at a queueing system.

$a(\tau) \triangleq$ number of arrivals at the queueing system in $[0, \tau]$

Next, we count the number of users leaving the system at its "exit" and let

$c(\tau) \triangleq$ number of service completions observed in $[0, \tau]$

If the system is empty at $t = 0$, the number of users in the system at the time $t = \tau$ is given by

$$N(\tau) = a(\tau) - c(\tau) \tag{4.6}$$

We can now use (4.6) to express the *total* amount of time, $I(\tau)$, spent by all users in the queueing system during the interval $[0, \tau]$. We have

$$I(\tau) = \int_0^\tau N(t)\,dt = \int_0^\tau [a(t) - c(t)]\,dt \tag{4.7}$$

Clearly, $I(\tau)$ represents the area between the functions $a(\tau)$ and $c(\tau)$, as illustrated in Figure 4.3.

The average number of users $\hat{N}(\tau)$ in the queueing system during the interval $[0, \tau]$ can now be obtained by dividing the total amount of time spent by all users in the queueing system, $I(\tau)$, by the time τ:

$$\hat{N}(\tau) = \frac{I(\tau)}{\tau} = \frac{I(\tau)}{a(\tau)} \frac{a(\tau)}{\tau} \tag{4.8}$$

We have written (4.8) in this form because both ratios on its right-hand side have very real physical meaning. The ratio $a(\tau)/\tau$ is simply the average number of arrivals per unit of time (the arrival rate) during the interval τ and can be indicated, given our earlier notation, as $\hat{\lambda}_\tau$. Similarly, $I(\tau)/a(\tau)$ is the average time spent by a user in the queueing system during the interval $[0, \tau]$ and can be indicated, given our earlier notation, as $\hat{W}_\tau$. We can then write

$$\hat{N}(\tau) = \hat{\lambda}_\tau \cdot \hat{W}_\tau \tag{4.9}$$

If we now let the length of the interval, τ, tend to infinity, it is clear from our earlier definitions of $\bar{L}$, λ, and $\bar{W}$ that these quantities represent the limits of $\hat{N}(\tau)$, $\hat{\lambda}_\tau$, and $\hat{W}_\tau$, respectively. So if the limits of the last two quantities ($\hat{\lambda}_\tau$ and $\hat{W}_\tau$) actually exist, the limit of $\hat{N}(\tau)$ also exists and from (4.9) we have the relationship

$$\bar{L} = \lambda \bar{W} \tag{4.10}$$

This is one of the best-known results of queueing theory and is referred to as *Little's formula*. Later in this chapter we shall explore the conditions under which the limits of $\hat{W}_\tau$ and of $\hat{N}(\tau)$ exist for many types of queueing systems.

A few important remarks are in order:

1. In deriving (4.10), we stationed ourselves at the "entrance" and "exits" of the queueing system. We could have performed exactly the same analysis if we had counted entries to the queueing system (as before) but exits from the *queue* (or, in other words, entries to the service facility). In that case we would have derived the result

$$\bar{L}_q = \lambda \bar{W}_q \tag{4.11}$$

where $\bar{L}_q$ and $\bar{W}_q$ are the average number and average stay in the *queue*, as defined earlier.

In a similar way, but by focusing now on the service facility itself (i.e., by counting entries and exits from the service facility), we could show that

$$\bar{L}_s = \lambda E[S] = \frac{\lambda}{\mu} \tag{4.12}$$

where $\bar{L}_s$ is the (steady-state) average number of users in the service facility. [Note that this result is independent of the number of servers, m. What matters on the right-hand side of (4.12) is the average amount of time a user spends in the facility, $E[S]$.]

2. In our discussion so far, we have never specified a queueing discipline. Thus, (4.10) and (4.11) hold irrespective of the method used to determine the order of entry to the service facility.[3] They also hold for the case where users belong to a number of distinct classes to which different levels of priority are assigned. Within each of the classes, (4.10) and (4.11) are valid.

3. The only condition that was placed on the arrival process at the queueing system is that the quantity $\lim_{\tau \to \infty} a(\tau)/\tau$, the long-term arrival rate, be finite. To appreciate the significance of this, consider a case in which the arrival rate, rather than being a constant, is a function of some parameter—say, of the total number of users present in the queueing system or of time. Then, we can still use (4.10) and (4.11), with a value of λ equal to *the long-term average of the rate at which users enter the system*. Similarly, for queueing systems with finite queue capacity, for which there is the possibility that some potential users will be turned away, we use (4.10) and (4.11) with λ equal to the average rate at which users *actually* join the queueing system. We shall see several examples of this type in subsequent sections.

A last relationship of importance which is always valid due to (4.1) is

$$\bar{W} = E[S] + \bar{W}_q = \frac{1}{\mu} + \bar{W}_q \tag{4.13}$$

Note that (4.10), (4.11), and (4.13) make it possible, with given λ and μ, to compute all four of the quantities $\bar{L}$, $\bar{L}_q$, $\bar{W}$, and $\bar{W}_q$ if any one of them can be determined.

Finally, it is worth remembering the following convenient argument (not "proof") that leads to (4.10) and (4.11). In the steady state, the average number of users that a random user finds in a FCFS "system" upon arrival should be equal to the number of users he or she leaves behind upon departure, with both of these numbers equal to $\bar{L}$ (or $\bar{L}_q$, depending on what the "system" is). But the average number of users left behind is simply the arrival rate λ times the average time a random user stays in the "system," $\bar{W}$ (or $\bar{W}_q$).

[3] See also Section 4.9.

4.5 FUNDAMENTAL QUEUEING MODEL

We turn now to the examination of a queueing model which we shall call the *fundamental birth-and-death model.* This model includes features that are quite general, and as a result, a rather extensive class of well-known and often-applied queueing systems can be viewed as simply special cases of the fundamental birth-and-death model.

The model assumes a queueing system with m ($m = 1, 2, 3, \ldots$) parallel identical servers and infinite system capacity, operating in the following fashion:

1. Whenever there are n users in the system (in queue plus in service) new users arrive at the system in a Poisson manner with a mean arrival rate of λ_n expected arrivals per unit time.
2. Whenever there are n users in the system, service completions occur in a Poisson manner with a mean service rate of μ_n per unit time.
3. The queueing system operates under a FCFS queue discipline.

It should be noted that μ_n here is defined as the combined rate of service of those servers which are busy when there are n users in the queueing system for any value of n. Note also that the rate of arrivals and the rate of service are permitted to depend on the number of users already in the system, a situation that is hardly unusual in actual life. We have already investigated two cases in Chapter 3 (see Sections 3.9.1 and 3.9.2) in which the arrival rate depended on the number of those already "in the system."

We can now go back to the review of the Poisson process (Section 2.12) and recognize that, if the number of users in the queueing system at a time t is given by $N(t)$, we can write the following conditional probabilities for our fundamental model:

$$P[N(t + \Delta t) = n + 1 \mid N(t) = n] = \lambda_n \Delta t + o(\Delta t) \tag{4.14}$$

$$P[N(t + \Delta t) = n - 1 \mid N(t) = n] = \mu_n \Delta t + o(\Delta t) \tag{4.15}$$

$$P[N(t + \Delta t) = n \mid N(t) = n] = 1 - \lambda_n \Delta t - \mu_n \Delta t - o(\Delta t) \tag{4.16}$$

$$P[N(t + \Delta t) = k \mid N(t) = n] = o(\Delta t) \quad \text{for } |k - n| > 1 \tag{4.17}$$

where $o(\Delta t)$, in the terminology of Chapter 2, is a "collection of terms that go to zero faster than $k \cdot \Delta t$ as Δt goes to zero (for any constant k)." *For a small* Δt, we can therefore ignore the $o(\Delta t)$ terms and state that "if the queueing system contains n users at time t, then at time $t + \Delta t$, it will contain either $n + 1$ users with probability $\lambda_n \Delta t$ or $n - 1$ users with probability $\mu_n \Delta t$, or n users with probability $1 - (\lambda_n + \mu_n) \Delta t$." This statement is illustrated schematically in Figure 4.4.

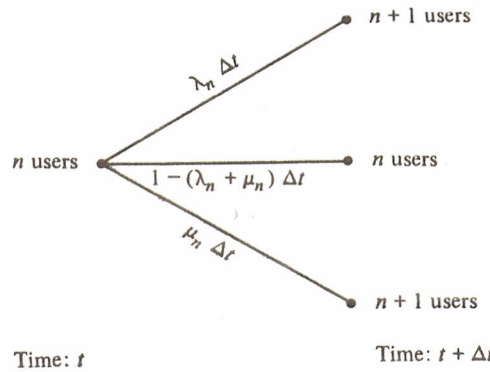

FIGURE 4.4 Probabilities of transitions for birth-and-death model in time Δt.

We can now proceed, as follows. Given (4.14)–(4.17), assuming a small Δt and using the notation $P_n(t) \triangleq P[N(t) = n]$, we can write

$$P_n(t + \Delta t) = P_{n+1}(t)\mu_{n+1} \Delta t + P_n(t)[1 - (\lambda_n + \mu_n) \Delta t] + P_{n-1}(t)\lambda_{n-1} \Delta t \tag{4.18}$$

Rearranging terms and dividing by Δt, we have

$$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = -(\lambda_n + \mu_n)P_n(t) + \mu_{n+1}P_{n+1}(t) + \lambda_{n-1}P_{n-1}(t) \tag{4.19}$$

Letting $\Delta t \to 0$, in (4.19) we obtain the differential equation

$$\frac{dP_n(t)}{dt} = -(\lambda_n + \mu_n)P_n(t) + \mu_{n+1}P_{n+1}(t) + \lambda_{n-1}P_{n-1}(t) \qquad n = 1, 2, 3, \ldots \tag{4.20}$$

This equation makes sense intuitively: it states that the rate of change of the probability of having exactly n users in the queueing system is equal to the probability of exactly $n + 1$ or $n - 1$ users in the system at time t multiplied, respectively, by the rate at which users leave or enter the system (with $n + 1$ and $n - 1$ users present, respectively) *minus* the probability that there are n users present at time t multiplied by the rate at which the number of users present can either increase (λ_n) or decrease (μ_n). Note the similarities between the interpretation of (4.20) and the interpretation of (2.55) in our discussion of the Poisson process in Chapter 2.

While (4.20) holds for $n = 1, 2, 3, \ldots$, we also require an equation for $n = 0$. By following a similar derivation or a similar logical argument as

above, we can write

$$\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \tag{4.21}$$

Equations (4.20) and (4.21) together define a set of differential equations, one for each possible value of n, for the queueing system analyzed here. Since, by assumption, the system capacity is infinite, so is the number of first-order differential equations.

A pictorial summary of the system of (4.20) and (4.21) is provided by Figure 4.5. The status of the queueing system is described by the state variable n (i.e., the total number of users in it). Thus, we shall say that the queueing system "is in state n" whenever there are n users in the system. Each state is represented by a circle in Figure 4.5 and the circles, in turn, are connected by directed links with the associated transition rate indicated on each link. Figure 4.5 is thus a typical *state transition diagram* for a queueing system. It is also clear from the diagram why our fundamental model is referred to as a birth-and-death model: in population applications of the model, the state n represents the "current population" and transitions out of state n can occur only to states $n + 1$ (a birth) and $n - 1$ (a death) at the rates λ_n and μ_n, respectively. Note also that, in this light the variants of the Poisson process described in Sections 3.9.1 and 3.9.2 (see Figures 3.35 and 3.36) can be considered "pure birth" processes.

Continuing with our analysis we can now examine the queueing system when it is in equilibrium (steady state). Under proper conditions, such an equilibrium will be reached after the system has been operating for some time. Equilibrium, in turn, implies that the state probabilities $P_n(t)$ eventually become independent of t and approach a set of constant values P_n, $n = 0, 1, 2, \ldots$, where[4]

$$P_n \triangleq \lim_{t \to \infty} P_n(t) = \text{steady-state probability that there are } n \text{ users in the queueing system}$$

Since $dP_n(t)/dt = 0$ under these circumstances, in the steady state, (4.20) and (4.21) are then transformed to

$$\lambda_0 P_0 = \mu_1 P_1 \tag{4.22}$$

$$(\lambda_n + \mu_n)P_n = \lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1} \qquad \text{for } n = 1, 2, 3, \ldots \tag{4.23}$$

The linear equations (4.24) and (4.25) are known as the *equilibrium equations* or the *balance equations* for our queueing system. Balance equations [as

[4]In Section 4.3 we used the more explicit notation $p_N(n)$ for the steady-state probability $P[N = n]$. From now on, for reasons of conciseness, we shall denote $P_N(n)$ by P_n. (Remember that N is the random variable indicating the total number of users in a queueing system under steady-state conditions.)

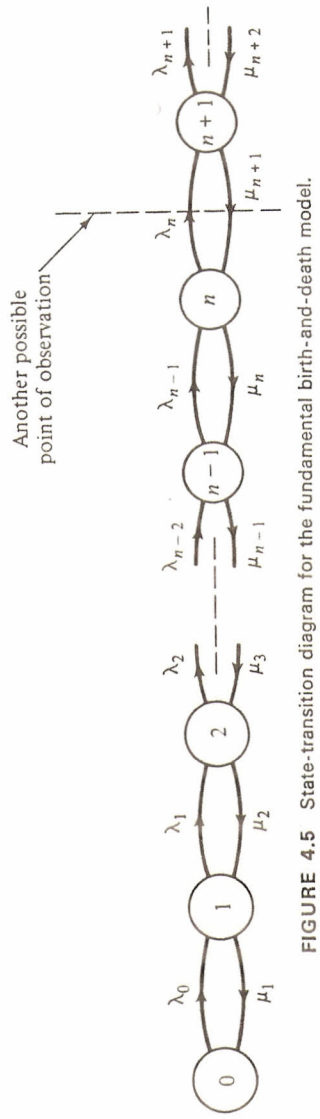

FIGURE 4.5 State-transition diagram for the fundamental birth-and-death model.

well as transition-rate differential equations such as (4.20) and (4.21)] can be written by inspection directly from the state-transition diagram of queueing systems. For the balance equation (4.23), for instance, the argument goes like this. For any time t when the system is in equilibrium, the probability of observing a transition out of state n in the next Δt must be equal to the probability of observing a transition into state n. The former quantity is given by P_n (= the probability that at time t the system is in state n) times $(\lambda_n + \mu_n)\,\Delta t$. Similarly, the probability of observing a transition into state n in the next Δt is given by $P_{n-1}\,\lambda_{n-1}\,\Delta t + P_{n+1}\,\mu_{n+1}\,\Delta t$.

Exercise 4.1 For birth-and-death queueing systems, another set of balance equations, even easier to solve than (4.22) and (4.23), can be obtained by (figuratively speaking) stationing ourselves at points such as those indicated by the dashed lines in Figure 4.5 and writing

$$\lambda_0 P_0 = \mu_1 P_1$$
$$\lambda_1 P_1 = \mu_2 P_2$$

and, in general,

$$\lambda_n P_n = \mu_{n+1} P_{n+1} \qquad \text{for } n = 0, 1, 2, 3, \ldots \tag{4.24}$$

a. Argue the validity of (4.24).

b. Derive (4.24) using (4.22) and (4.23).

4.5.1 Solving the Balance Equations

We can now proceed to solve the balance equations expressing all steady-state probabilities P_n, $n = 0, 1, 2, \ldots$ in terms of one of them and then taking advantage of the fact that

$$\sum_{n=0}^{\infty} P_n = 1 \tag{4.25}$$

It is common practice in queueing theory to express $P_1, P_2, P_3, \ldots$ in terms of P_0, the steady-state probability of an empty system. Working with (4.22) and (4.23) or, equivalently (and preferably) with (4.24), we have

$$P_1 = \frac{\lambda_0}{\mu_1} P_0 \tag{4.26}$$

$$P_2 = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0 \tag{4.27}$$

and, in general,

$$P_n = \frac{\lambda_{n-1} \cdot \lambda_{n-2} \cdot \cdots \cdot \lambda_1 \cdot \lambda_0}{\mu_n \cdot \mu_{n-1} \cdot \cdots \cdot \mu_2 \cdot \mu_1} P_0 \tag{4.28}$$

or, defining the coefficient of P_0 in (4.28) as the quantity K_n, we have

$$P_n = K_n P_0 \qquad \text{for } n = 1, 2, 3, \ldots$$

Going now back to (4.25),

$$\sum_{n=0}^{\infty} P_n = \left(1 + \sum_{n=1}^{\infty} K_n\right) P_0 = 1 \tag{4.29}$$

or

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} K_n} \tag{4.30}$$

It follows that the system can reach steady state only if $\sum_{n=1}^{\infty} K_n < \infty$. For, otherwise, $P_0 = P_1 = P_2 = \ldots = 0$ (i.e., the number of users in the system never "stabilizes").

Assuming that the system does reach steady state, the probabilities P_n, $n = 0, 1, 2, \ldots$, are now given by the fundamental expressions (4.30) and (4.28), and other quantities of interest can be computed from these probabilities. For instance,

$$\bar{L} = \sum_{n=0}^{\infty} n P_n \tag{4.31}$$

and $\bar{L}_q$, $\bar{W}$, and $\bar{W}_q$ can then be obtained from (4.10), (4.11), and (4.13).

Exercise 4.2 Argue that the value of λ that should be used with Little's equations, (4.10) and (4.11) is in this case,

$$\lambda = \sum_{n=0}^{\infty} \lambda_n P_n \tag{4.32}$$

4.6 CENTER FOR EMERGENCY CALLS: QUEUEING SYSTEMS OF THE BIRTH-AND-DEATH TYPE

Equipped with the results that we have derived for our fundamental birth-and-death model, we shall now review a problem whose many variations will help illustrate several of the best-known and most widely used models of queueing systems.

Many cities have by now instituted the use of the telephone number 911 for all types of emergency calls. At the "other end" of the 911 number there is usually a center for emergency calls employing a number of trained operators. The sophistication of equipment and operational setup in these centers varies widely from city to city. Some cities (e.g., New York City) employ

quite elaborate schemes for screening calls and determining their priorities, schemes backed up by special-purpose computers and communications equipment. Other centers consist of little more than a switchboard and a number of telephone operators who either process telephone calls themselves (in cooperation with a number of dispatchers) or transfer calls to the most appropriate city department (e.g., fire department, emergency medical services department, etc.). It is interesting to compare, at least in an approximate way, the characteristics of these centers as a function of different levels of manpower and under various organizational schemes. Queueing theory offers us a good opportunity to do so.

Throughout the following discussion it will be assumed that the arrival of calls at a center constitutes a Poisson process (whose mean rate may vary). This assumption is reasonable, with the possible exception of the occurrence of major incidents which can be expected to trigger bursts of telephone calls—all reporting the same event and its repercussions.

4.6.1 Case 1: One Operator, Infinite Number of Lines

Consider first a case in which a single operator answers all calls and in which the number of telephone lines leading into the switchboard is large enough so that the system never runs out of lines. If the operator is busy when a call arrives, the caller hears a taped message to "please wait" and his or her call joins the queue of 0, 1, 2, . . . more callers already waiting for the single operator. We assume that:

1. No caller ever gets sufficiently discouraged from waiting to hang up.
2. An electronic device sequences calls so that they are answered in a FCFS order.
3. The pdf for call interarrival times is negative exponential with mean $1/\lambda$ and the pdf for service times (duration of telephone conversations) is negative exponential with mean $1/\mu$.

What has been described is a queueing system of the $M/M/1$ type with FCFS service and an infinite system capacity. The latter is due to the large number of lines, which theoretically are sufficient so that there is always an open line to the switchboard. A state-transition diagram for this system is shown in Figure 4.6.

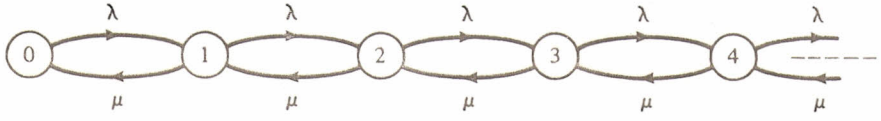

FIGURE 4.6 State-transition diagram for a $M/M/1$ queueing system with infinite system capacity.

In terms of our fundamental model, we then have

$$\lambda_n = \lambda \qquad \text{for } n = 0, 1, 2, \ldots$$
$$\mu_n = \mu \qquad \text{for } n = 1, 2, 3, \ldots$$

and substituting into (4.28) and (4.30)—and recognizing the presence of a geometric series for $\rho\ (= \lambda/\mu)$—we find that

$$P_0 = 1 - \rho \tag{4.33}$$
$$P_n = \rho^n(1 - \rho) \qquad \text{for } n = 1, 2, 3, \ldots \tag{4.34}$$

with the condition for steady state being that the geometric series converges, that is, that

$$\rho = \frac{\lambda}{\mu} < 1 \tag{4.35}$$

The condition (4.35) makes sense intuitively. If $\rho > 1$ (i.e., if $\lambda > \mu$), the average rate of arrivals at the queueing system exceeds the rate at which the server-operator can service calls. Thus, the longer the system operates, the longer the queue tends to become and no steady state is ever reached.

It is much less obvious why steady state is not reached for $\rho = 1$. A possible way for explaining this is to argue that the longer the queue grows in this case, the more unlikely it is that it will ever appreciably decrease again, since the service rate just matches the arrival rate.

Numerous other quantities can now be computed for the $M/M/1$ system. For instance (the algebra here is quite interesting),

$$\bar{L} = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n(1 - \rho)\rho^n = (1 - \rho)\rho \sum_{n=1}^{\infty} n\rho^{n-1}$$
$$= (1 - \rho)\rho \frac{d}{d\rho}\left(\sum_{n=0}^{\infty} \rho^n\right) = (1 - \rho)\rho \frac{d}{d\rho}\left(\frac{1}{1 - \rho}\right) = \frac{(1 - \rho)\rho}{(1 - \rho)^2}$$

or

$$\bar{L} = \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda} \tag{4.36}$$

Similarly,

$$\bar{L}_q = \sum_{n=1}^{\infty} (n - 1)P_n = \bar{L} - (1 - P_0) = \frac{\rho^2}{1 - \rho} = \frac{\lambda^2}{\mu(\mu - \lambda)} \tag{4.37}$$

Also,

$$\bar{W} = \frac{1}{\mu}P_0 + \frac{2}{\mu}P_1 + \frac{3}{\mu}P_2 + \cdots = \sum_{n=0}^{\infty} \frac{n + 1}{\mu} P_n$$
$$= \frac{1}{\mu(1 - \rho)} = \frac{1}{\mu - \lambda} \tag{4.38}$$

and

$$\bar{W}_q = \sum_{n=0}^{\infty} \frac{n}{\mu} P_n = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)} \tag{4.39}$$

[Of course, (4.37)–(4.39) could have been obtained from (4.36) by just using (4.10), (4.11), and (4.13).]

In addition to the expected total system time, $\bar{W}$, it is actually possible to derive, for a $M/M/1$ queue, the pdf for the system occupancy time, W, under steady-state conditions. To do this, let A_n be the event that a random call arrives to find n other calls already at the switchboard and thus becomes the $(n+1)$th call in the system. (Clearly, $P\{A_n\} = P_n$.) The total time that caller spends in the system (in queue and in service) is the sum of $n+1$ independent and negative exponentially distributed[5] random variables each with mean $1/\mu$. Thus, W, for any given n, has the pdf of an Erlang random variable of order $n+1$. So, we have

$$f_{W|A_n}(w \mid A_n) = \frac{\mu^{n+1} w^n}{n!} e^{-\mu w} \qquad \text{for } n = 0, 1, 2, \ldots$$

and[6]

$$f_W(w) = \sum_{n=0}^{\infty} P_n \cdot f_{W|A_n}(w \mid A_n) = \sum_{n=0}^{\infty} (1-\rho)\rho^n \frac{\mu^{n+1} w^n}{n!} e^{-\mu w}$$
$$= (1-\rho)\mu e^{-\mu w} \cdot e^{\mu \rho w}$$

or

$$f_W(w) = (1-\rho)\mu e^{-(1-\rho)\mu w} = (\mu - \lambda) e^{-(\mu-\lambda)w} \qquad \text{for } w \geq 0 \tag{4.40}$$

Thus, somewhat surprisingly, the system occupancy time is exponentially distributed. Its expected value, $\bar{W}$, is equal to $[\mu(1-\rho)]^{-1}$, which, of course, is the result we have already obtained in (4.38).

In a quite similar way, one can derive the pdf for the time spent waiting in the queue, W_q. This pdf is a mixed one, having an impulse at $W_q = 0$. The reason is that with probability $P_0 = 1 - \rho$, the waiting time will be equal to 0. It is thus more convenient to write the cumulative distribution for W_q:

$$F_{W_q}(t) = \begin{cases} 1-\rho & \text{for } t = 0 \\ 1 - \rho e^{-\mu(1-\rho)t} & \text{for } t > 0 \end{cases} \tag{4.41}$$

[5]The remaining service time of the caller who occupies the operator at the time of arrival of the caller of interest is also distributed as a negative exponential random variable because of the "no-memory" property of the negative exponential pdf.

[6]Remember that $\sum_{n=0}^{\infty} x^n/n! = e^x$.

Exercise 4.3 Show that the variance of the total number of callers in the system is

$$\sigma_N^2 = \sum_{n=0}^{\infty} (n - \bar{L})^2 P_n = \frac{\rho}{(1-\rho)^2} \tag{4.42}$$

Thus, as $\rho \to 1$, not only does the expected number of callers, $\bar{L}$, in the queueing system grow in proportion to $(1-\rho)^{-1}$ but also the variance grows as $(1-\rho)^{-2}$ (i.e., even faster).

Exercise 4.4 Finally, our last result refers to the average duration of a *busy period* for the $M/M/1$ system. A busy period is defined to begin with the arrival of a call while the system switchboard is completely empty and to end when the switchboard next becomes free once again. Argue that in steady state,

$$E[B] \triangleq E[\text{length of a busy period}] = \frac{1}{\mu - \lambda} \tag{4.43}$$

4.6.2 Case 2: m Operators, Infinite Number of Lines

Suppose now that, while keeping everything else in the emergency call center exactly the same as before, the number of telephone operators is increased to m (>1). The service time pdf's associated with *each operator* are identical and negative exponential with parameter μ. When all operators are busy, the next call in line is assigned to the first operator to become free, while when two or more operators are free, the next incoming call is assigned to an operator in some arbitrary way.

The state-transition diagram for this case is shown in Figure 4.7. In terms of the queueing system code, this is a $M/M/m$ system with infinite queue capacity and FCFS service. With respect to our fundamental model,

$$\lambda_n = \lambda \qquad n = 0, 1, 2, \ldots$$
$$\mu_n = n\mu \qquad n = 1, 2, 3, \ldots, m-1$$
$$\mu_n = m\mu \qquad n = m, m+1, m+2, \ldots$$

Then, from (4.28),

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & \text{for } n = 0, 1, \ldots, m-1 \\ \dfrac{(\lambda/\mu)^n}{m^{n-m} \cdot m!} P_0 & \text{for } n = m, m+1, m+2, \ldots \end{cases} \tag{4.44}$$

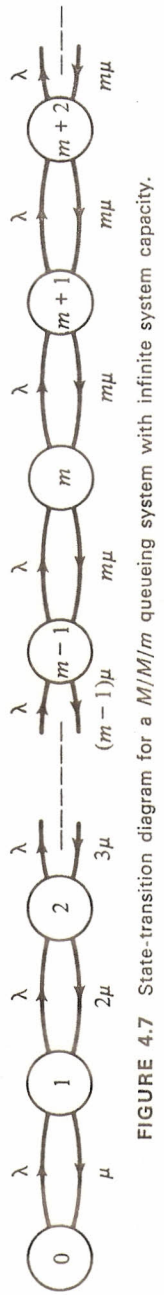

FIGURE 4.7 State-transition diagram for a $M/M/m$ queueing system with infinite system capacity.

and, substituting in (4.25),

$$1 = \sum_{n=0}^{\infty} P_n = \left[\sum_{n=0}^{m-1} \frac{(\lambda/\mu)^n}{n!} + \sum_{n=m}^{\infty} \frac{(\lambda/\mu)^n}{m^{n-m} \cdot m!}\right] P_0$$
$$= \left[\sum_{n=0}^{m-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^m}{m!} \sum_{n=0}^{\infty} \left(\frac{\lambda}{m\mu}\right)^n\right] P_0 \tag{4.45}$$

Using the geometric series expression for the last sum in the brackets (assuming that $\lambda/m\mu < 1$), we finally have from (4.45)

$$P_0 = \left[\sum_{n=0}^{m-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^m}{m!} \frac{1}{1 - (\lambda/m\mu)}\right]^{-1} \tag{4.46}$$

with the condition for steady state being

$$\frac{\lambda}{m\mu} < 1 \tag{4.47}$$

Expressions for other quantities of interest can now be derived using the steady-state probabilities, P_n.

Limiting case: Infinite number of servers. The limiting extension of case 2 is when the number of servers m is (countably) infinite. In such a situation no user of the queueing system will ever have to wait in line. Since in this case we have

$$\lambda_n = \lambda \qquad \text{for } n = 0, 1, 2, \ldots \tag{4.48a}$$
$$\mu_n = n\mu \qquad \text{for } n = 1, 2, 3, \ldots \tag{4.48b}$$

it follows from (4.28) that

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_0 \qquad \text{for } n = 1, 2, 3, \ldots \tag{4.49}$$

and, using (4.25),

$$1 = \sum_{n=0}^{\infty} P_n = P_0 \sum_{n=0}^{\infty} \frac{(\lambda/\mu)^n}{n!} = P_0 e^{\lambda/\mu}$$

With (4.49) we thus conclude that

$$P_n = \frac{(\lambda/\mu)^n e^{-\lambda/\mu}}{n!} \qquad \text{for } n = 0, 1, 2, \ldots \tag{4.50}$$

This is a remarkable result, stating that the steady-state probability distribution for the number of users present (and, consequently, for the number of busy servers as well) in a $M/M/\infty$ system is Poisson with parameter λ/μ.

It follows, of course, that $\bar{L} = E[N] = \lambda/\mu$, $\bar{W} = 1/\mu$, and that $\bar{W}_q = \bar{L}_q = 0$. Note also that steady state is inevitably reached in this case, since there are always sufficiently many servers to assure that the service rate will eventually exceed the rate of arrivals [see (4.48)].

Although one may argue that there are not many systems around with an infinite number of servers, the model of this section is still very useful in numerous applications in which there is only a very low probability that all the servers in a system with many parallel, identical servers will be busy simultaneously. The Poisson distribution result for the steady-state probabilities [(4.50)] that we derived can then be used to obtain good approximations of occupancy-related statistics for the system in question. We shall return to this type of approximation in our subsequent discussion of the $M/G/\infty$ queueing system (Section 4.8).

4.6.3 Case 3: One Operator, Finite Number of Lines

Let us now continue with our emergency call center example and consider a situation identical to that of case 1 with one exception. Rather than an infinite number, there is now only a finite number of lines, K, into the switchboard. Furthermore, the analysis will be performed under the assumption that a caller who calls a 911-type number and gets a busy signal becomes discouraged and does not try again. We shall discuss the implications of this assumption at the end of this section.

The state-transition diagram for case 3 is shown in Figure 4.8. This is a $M/M/1$ system with *finite* system capacity equal to K. With respect to our fundamental birth-and-death model,

$$\lambda_n = \begin{cases} \lambda & \text{for } n = 0, 1, 2, \ldots, K-1 \\ 0 & \text{otherwise} \end{cases} \tag{4.51a}$$

$$\mu_n = \begin{cases} \mu & \text{for } n = 1, 2, \ldots, K \\ 0 & \text{otherwise} \end{cases} \tag{4.51b}$$

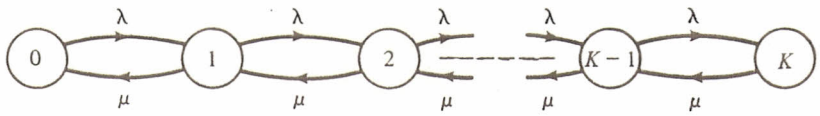

FIGURE 4.8 State-transition diagram for a $M/M/1$ queueing system with system capacity equal to K.

It follows from (4.28) that

$$P_n = \begin{cases} \left(\frac{\lambda}{\mu}\right)^n P_0 = \rho^n P_0 & \text{for } n = 0, 1, 2, \ldots, K \\ 0 & \text{otherwise} \end{cases} \tag{4.52}$$

and therefore,

$$1 = \sum_{n=0}^{K} P_n = P_0(1 + \rho + \rho^2 + \cdots + \rho^K) = P_0 \frac{1 - \rho^{K+1}}{1 - \rho} \tag{4.53}$$

As a result of (4.52) and (4.53), we finally obtain for case 3:

$$P_n = \begin{cases} \frac{\rho^n(1-\rho)}{1-\rho^{K+1}} & \text{for } n = 0, 1, 2, \ldots, K \\ 0 & \text{otherwise} \end{cases} \tag{4.54}$$

Note that for $\rho < 1$, (4.54) reduces to (4.34), as it should, as $K \to \infty$. Knowing the steady-state probabilities, we can now obtain expressions for $\bar{L}$, $\bar{W}$, $\bar{L}_q$, and $\bar{W}_q$, one of which is listed in Table 4-1. Aside, however, from the specific form of the results obtained, there are two points which are worth

TABLE 4-1 Summary of steady-state results for some simple queueing systems

1. $M/M/1$ (system capacity infinite)
   See section 4.6.1
2. $M/M/m$ (system capacity infinite)
   See section 4.6.2. In addition:
   $$\bar{L}_q = \frac{P_0(\lambda/\mu)^m(\lambda/m\mu)}{m!(1-\lambda/m\mu)^2}$$
3. $M/M/1$ (system capacity K)
   See section 4.6.3. In addition:
   $$\bar{L}_q = \frac{\rho}{1-\rho} - \frac{\rho(1+K\rho^K)}{1-\rho^{K+1}}$$
4. $M/M/m$ (system capacity K)
   $$P_0 = \left[\sum_{n=0}^{m} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^m}{m!} \sum_{n=m+1}^{K} \left(\frac{\lambda}{m\mu}\right)^{n-m}\right]^{-1} \qquad \left(\text{for } \frac{\lambda}{m\mu} < 1\right)$$
   $$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0 & \text{for } n = 0, 1, 2, \ldots, m \\ \frac{(\lambda/\mu)^n}{m!\, m^{n-m}} P_0 & \text{for } n = m+1, m+2, \ldots, K \end{cases}$$
   $$\bar{L}_q = \frac{P_0(\lambda/\mu)^m(\lambda/m\mu)}{m!(1-\lambda/m\mu)^2}\left[1 - \left(\frac{\lambda}{m\mu}\right)^{K-m} - (K-m)\left(\frac{\lambda}{m\mu}\right)^{K-m}\left(1 - \frac{\lambda}{m\mu}\right)\right]$$

remembering about the $M/M/1$ system with finite system capacity. Moreover, these points are valid for finite capacity systems in general:

1. Steady-state conditions will be reached in any event, irrespective of the value of ρ. Because the number of states in the system is finite, there is an upper limit on how long the queue can get. Even when the arrival rate is much larger than the service rate ($\rho \gg 1$), steady

state will be reached: the queue will simply be full most of the time. In such an event a large fraction of the potential facility users will be turned away. That fraction is equal to P_K, the probability that the queue is saturated. Note also that for $\rho = 1$, $P_n = 1/(K+1)$ for all values of n (i.e., all states are equally likely).

2. The fundamental relationships $\bar{L} = \lambda\bar{W}$ and $\bar{L}_q = \lambda\bar{W}_q$ still hold but, as suggested by (4.32), in revised form. Since a fraction P_K of potential users of the facility are turned away, the *actual arrival rate* at the queue is equal to $\lambda' \triangleq \lambda(1 - P_K)$. Thus, the relationships (4.10) and (4.11) are now revised to the form

$$\bar{L} = \lambda'\bar{W} \tag{4.55}$$

and

$$\bar{L}_q = \lambda'\bar{W}_q \tag{4.56}$$

where λ' is as defined above. The third fundamental relationship, $\bar{W} = \bar{W}_q + 1/\mu$, still holds.

In practice it is rather unlikely that an emergency caller who gets a busy signal will refrain from calling the emergency center again. (However, this may be true for nonemergency callers, and it would definitely be true for queueing systems that offer routine types of services that can be obtained with ease at other queueing systems, as well.) If some callers persist in calling the emergency number, the resulting situation is an intermediate one between case 3 and case 1. In fact, in the extreme case when no caller ever becomes discouraged and they all keep trying continually to get a free line, an infinite capacity system will again result. However, during periods of congestion, we now have two queues: a "visible" one consisting of the K callers who have already obtained access to the switchboard, and an "invisible" one consisting of all those trying to obtain such access. In addition, while the former queue is operating on a FCFS basis (because of the existence of the call-ordering "electronic device"), access to the switchboard from the invisible queue is of the SIRO type.

Finally, it should be clear that for system design purposes, the probability P_K of a full system is often the most important design parameter. For, when P_K is negligibly small, a finite-capacity system operates, for all practical purposes, like a system with infinite capacity.

4.6.4 Case 4: m Operators, Finite Number of Lines

Case 4 involves a finite number, m, of operators and a finite number, K, of lines. In many ways, this is probably the most general and appropriate model for a simple emergency call center. The "design parameters" then

involve the determination of the right number of operators and lines so that a combination of objectives will be achieved. These objectives might be in the form of specifications, for instance, of an upper limit on (1) the probability that a caller receives a busy signal, P_K; and (2) some other level-of-service indicator, such as the expected system occupancy time, $\bar{W}$, for a random *accepted* caller. An interesting analysis of this type has been reported for the 911 emergency call center in New York City [LARS 12a] (see also Chapter 8).

A state-transition diagram for a $M/M/m$ system with a finite system capacity is shown in Figure 4.9, and expressions for some quantities related to this case are listed in Table 4-1. The algebra involved gets quite tedious for the general case, but numerical applications are straightforward and can be performed easily with a hand calculator.

It should finally be emphasized that, unless otherwise stated, the implicit assumption in finite-system-capacity queueing models is that prospective users who find a full system (with probability P_K) are permanently lost to the system.

Special case: $K = m$ (Erlang's loss formula). A special instance of the case 4 system is when the capacity of the system, K, is equal to the number of servers, m. Such would be the case if there were one telephone line for each operator. Obviously, in such a system, there is no waiting space at all and users who, on arrival, find all servers busy are simply turned away. Historically, this was one of the first queueing systems ever to be investigated in depth. This was done by A. K. Erlang of Denmark (generally considered to be the "father" of queueing theory) during the first decade of this century.

The interesting quantities for this case can be obtained by setting $K = m$ in the expressions obtained for case 4 above (see Table 4-1). However, it is just as easy to work directly with the balance equations and obtain

$$P_n = \frac{(\lambda/\mu)^n/n!}{\sum_{i=0}^{m} (\lambda/\mu)^i/i!} \qquad n = 0, 1, 2, \ldots, m \tag{4.57}$$

In particular, the probability of a full system, P_m (i.e., the probability that an arriving user will find a full system and be turned away) is widely known as *Erlang's loss formula* and has been extensively tabulated for different values of the ratio λ/μ (which is not restricted as to magnitude) and of the number of servers m. Erlang's loss formula has also been used widely in applications of queueing theory to urban service systems (see also Section 4.8).

It should be clear that the $M/M/\infty$ queueing system can also be viewed as a special case of $M/M/m$ with no waiting space. In fact, by letting m go to infinity in (4.57) we obtain expression (4.50) for the steady-state probabilities of the $M/M/\infty$ queueing system.

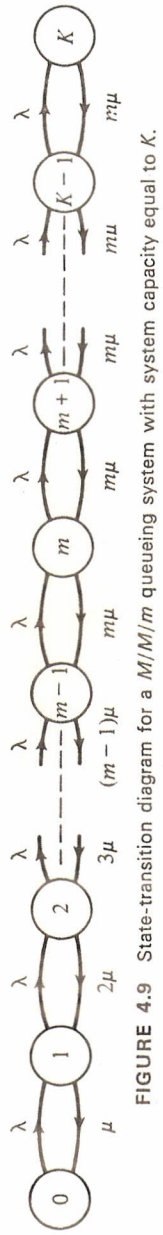

FIGURE 4.9 State-transition diagram for a $M/M/m$ queueing system with system capacity equal to K.

We shall also anticipate here an interesting result that will be presented later in this chapter. It turns out that (4.57) holds for any service-time distribution, that is, for $M/G/m$ queueing systems with no waiting space!

4.6.5 Extensions and Variations

Further extensions and variations of the many cases discussed in this section can be developed by permitting the arrival rates and/or the service rates to take the form

$$\lambda_n = c_n\lambda \qquad n = 0, 1, 2, \ldots \tag{4.58a}$$

$$\mu_n = d_n\mu \qquad n = 1, 2, 3, \ldots \tag{4.58b}$$

where λ and μ are constants and c_n and d_n are coefficients that depend on the state of the queueing system. (Usually, $c_0 = 1$ and $d_1 = 1$, so that λ reflects the rate of demand arrivals when the queueing system is empty and μ the rate of service when there is only one user in the queueing system.)

Through (4.58a), one can take into account often-observed phenomena such as reneging and balking. *Balking* refers to cases in which a prospective user of the queueing system decides, upon arrival at the system and observation of its state, not to wait for its use but (perhaps) to go elsewhere. *Reneging* is the phenomenon in which users who have *already* joined the queue become discouraged after a while and leave without obtaining service.

Similarly, by an appropriate choice of a functional form for d_n, one can account for such phenomena as the often-observed "speed up" of service by human operators whenever queues grow very long. It is easy to imagine how reneging, balking, service speed up or, even, service slowdowns could occur in connection with our emergency center example.

Commonly used forms of c_n include $c_n = (n + 1)^{-a}$ and $d_n = n^b$, where a and b are positive constants. Note how by adjusting a and/or b one can model several types of user behavior. These can be applied with single- or multi-server systems whose capacity is finite or infinite. Although it is usually impossible to obtain closed-form expressions for such quantities as P_n, $\bar{L}$, $\bar{W}$, and so on, for these systems, it is often relatively easy to solve *numerically* the balance equations and tabulate the numerical results (see also Problem 4.2). Such tabulations have been published by several researchers.

4.7 SPATIALLY DISTRIBUTED QUEUES AND THE *M/G/1* QUEUEING SYSTEM

As we have already mentioned at the beginning of this chapter, many queueing systems in the urban environment do not fit neatly into the classical mold of a physically stationary server where prospective "customers" arrive and

queue up until they receive service. For most of the emergency urban services—where "arrivals of customers" are perceived through telephone calls (or other means of telecommunication) from various locations in a city—the only place at which a "queue" can be identified is on the emergency system's dispatcher's desk (or in a computer's memory), where records indicate the time, origin, and nature of a succession of requests for service. The server, then, be it a person or a vehicle, must travel to the location of these incidents to provide the required service. In such cases we have a *spatially distributed queue*.

In this section we shall discuss the simplest possible type of spatially distributed queue in which a single server has sole responsibility for a given district. This discussion will also motivate our derivation of some important results for $M/G/1$ queueing systems.

Example 1: Ambulance Service to an Emergency Medical Facility

Consider the case pictured in Figure 4.10: an emergency medical facility (EMF) is located at the center (the point where the diagonals intersect) of a rectangular district with dimensions $X_0 \times Y_0$ miles. The EMF has a single ambulance vehicle associated with it, which is dispatched to emergency patients and transports them to the EMF. The ambulance, when idle, is always located at the EMF.

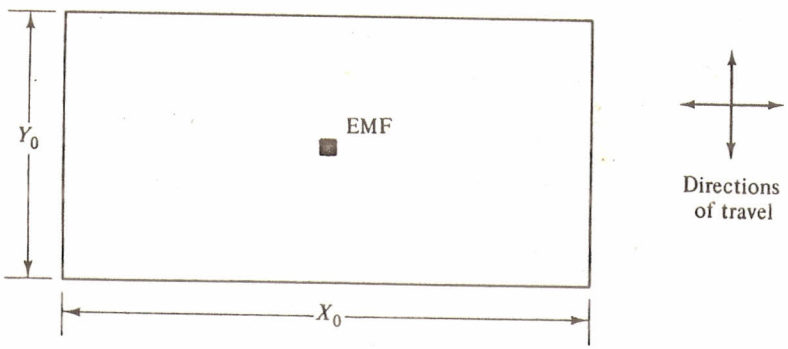

FIGURE 4.10 Rectangular district with an emergency medical facility located at its center.

Directions of travel are parallel to the boundaries of the district and the effective travel speeds are v_x and v_y (constants) in the x and y directions, as shown in Figure 4.10.

In this district, incidents (each corresponding to one emergency patient) occur as a Poisson process in time at a rate of λ per hour. Incident locations are independent of each other and are *uniformly* distributed over the rectangular area. Every time an incident occurs, the ambulance, if at the EMF, is immediately dispatched to the location of the call, picks up the patient, and

returns to the EMF. If the ambulance is away, calls queue up and are processed in a FIFO order. We shall assume for now that no calls are ever lost, no matter how long they have to wait. We shall also assume that the pdf for the time, Z, that the ambulance spends at the location of each call for service is known and is given by $f_Z(z)$ with an expectation $\bar{Z}$ and a variance σ_Z^2.

The service time, S, in this system clearly consists of a "travel-time" component and a "time on the scene" component. Assuming that effective travel speeds are identical on both the EMF-to-incident and the incident-to-EMF portions of each trip, we have

$$S = 2(T_x + T_y) + Z \tag{4.59}$$

where $T_x = D_x/v_x$ and $T_y = D_y/v_y$, and D_x and D_y are defined as the distances along the x and the y axes, respectively, from (to) the EMF to (from) the location of an incident.

With the techniques presented in Chapter 3, it is an easy matter to obtain the expected value of S, which for consistency with our queueing theory notation we shall denote as $1/\mu$:

$$\frac{1}{\mu} = E[S] = 2\left(\frac{X_0}{4v_x} + \frac{Y_0}{4v_y}\right) + \bar{Z} = \frac{1}{2}\left(\frac{X_0}{v_x} + \frac{Y_0}{v_y}\right) + \bar{Z} \tag{4.60}$$

Similarly, assuming that time at the scene of a call is independent of travel time (a quite reasonable assumption in this case), we have for the variance of S,

$$\sigma_S^2 = 4(\sigma_{T_x}^2 + \sigma_{T_y}^2) + \sigma_Z^2 = \frac{1}{12}\left(\frac{X_0^2}{v_x^2} + \frac{Y_0^2}{v_y^2}\right) + \sigma_Z^2 \tag{4.61}$$

If $f_Z(z)$, as we have already assumed, is known, it is also possible, at least in principle, to obtain an expression for $f_S(s)$, the service-time pdf. The derivation, however, even for the simple situation described here, promises to be tedious and time-consuming and will be omitted. It suffices to note that $f_S(s)$ cannot be a negative exponential pdf [unless $f_Z(z)$ is negative exponential *and* the expected time on the scene is much larger than the average round-trip time, in which case it can be argued that $1/\mu \approx \bar{Z}$ and, therefore, that $f_S(s)$ is approximately negative exponential as well[7]]. It is therefore clear that results derived for the single-server queueing systems we have seen so far ($M/M/1$) do not apply in this case, since the service-time distribution at hand is a "general" one.

We shall now proceed to derive several important results for this $M/G/1$ queueing system. Interestingly, as we shall see, most of the results require no knowledge of the service-time distribution, $f_S(s)$, other than its mean, $1/\mu$, and variance σ_S^2.

[7]It turns out that this is often the case with some important urban emergency services, such as police and most emergency repair services. This makes possible the derivation of many powerful results with respect to these services (see Chapter 5).

Recall first that the current state of $M/M/1$ (or $M/M/m$) queueing systems is fully described by a single item of information, the number of users (i.e., of calls in our EMF example) currently in the system. Knowledge of this number is sufficient to describe all the past history of the queueing system, as far as the future is concerned. For instance, if it is known that at some time instant, t, there are exactly n (> 0) calls in a $M/M/1$ system (with one call receiving service and the other $n - 1$ calls in the waiting line), then we can immediately state that the probability that in the next Δt a service will be completed is equal to $\mu\,\Delta t$—independently of what else has happened in the past at that $M/M/1$ system. For $M/G/1$ systems, however, this probability also depends on how long ago service began to the call that is currently receiving service. Thus, a complete description of the current state of a $M/G/1$ system requires, in general, specification of the values of two random variables, the number of calls currently in the system, and the time since the current service began, the latter of which is a continuous random variable. These complications make the mathematical analysis of $M/G/1$ systems more difficult than that of $M/M/1$ or $M/M/m$ systems.

Of the several different approaches that have been developed, the simplest one uses the trick of focusing attention on certain specific instants in time, known as *epochs*, when knowledge of only the number of calls currently in the queueing system is sufficient to specify its current state. Those instants of time are the times of completion of a service by the server (i.e., the instants after the ambulance has returned to the EMF and delivered a patient, in the case of our example).

Let us then indicate these time instants as $t_1, t_2, t_3, \ldots$ with t_i representing the instant when service to the ith patient to be transported to the EMF (beginning with some arbitrary time $t = 0$) is completed. A specific example for a hypothetical $M/G/1$ queueing system is shown in Figure 4.11.

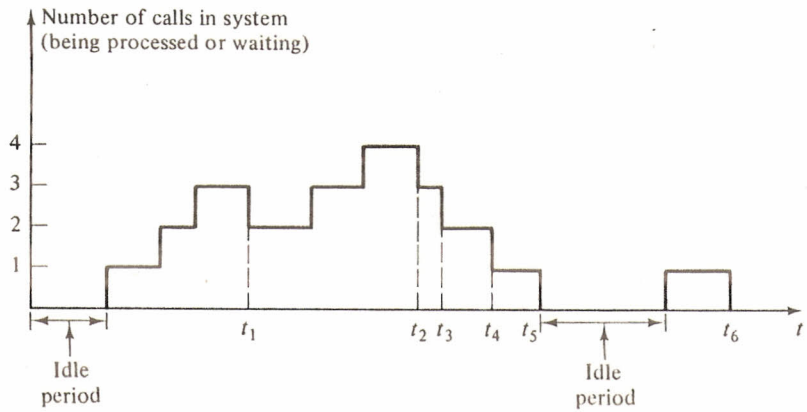

FIGURE 4.11 Possible time history for the $M/G/1$ system of our example with six epochs, $t_1, t_2, \ldots t_6$.

We can then define:

N = number of calls in the queueing system (i.e., the number of patients/incidents waiting for service) just after the instant t_{i-1} when service to the $(i - 1)$th patient is completed

R = number of *new* calls that arrive at the queueing system during the service time of the next patient to receive service (patient i)

N' = number of calls in the queueing system just after t_i, the instant of completion of service to patient i

Note that, by definition, R includes only those calls that arrive at the queueing system *after* service to the next patient (patient i) has started. This is important for the cases when, upon completion of a service, the queueing system is left empty (i.e., no more calls left to serve). For instance, the value of R for the time interval between t_5 and t_6 is 0 (*not* 1) for the sample case shown in Figure 4.11.

The following relationship exists between the random variables N, R, and N':

$$N' = \begin{cases} N + R - 1 & \text{if } N > 0 \\ R & \text{if } N = 0 \end{cases} \tag{4.62}$$

The probability α_r that exactly r calls arrive during a service time is given by

$$\alpha_r = P\{\text{number of new arrivals during a service time} = r\} = p_R(r)$$
$$= \int_0^\infty \frac{(\lambda t)^r e^{-\lambda t}}{r!} f_S(t)\,dt \qquad \text{for } r = 0, 1, 2, \ldots \tag{4.63}$$

where, as above, $f_S(s)$ represents the pdf for the service time. For any given pdf $f_S(s)$, it is then possible to determine the probabilities α_r.

Exercise 4.5 How is (4.63) justified?

Hint: *Given* that a service lasted exactly a time t, what is the probability that r new calls arrived during that service time?

For $r = 0, 1, 2, \ldots$, we have

$$P\{n + r - 1 \text{ users present at } t_{k+1} \mid n \text{ users present at } t_k\} = \alpha_r \qquad \text{for } n > 0 \tag{4.64}$$

$$P\{r \text{ users present at } t_{k+1} \mid 0 \text{ users present at } t_k\} = \alpha_r \tag{4.65}$$

So (4.64) and (4.65) give the state-transition probabilities for successive epochs for the $M/G/1$ system. The state-transition diagram for our $M/G/1$ system, *at the epochs* is now shown in Figure 4.12. A state is defined by "the

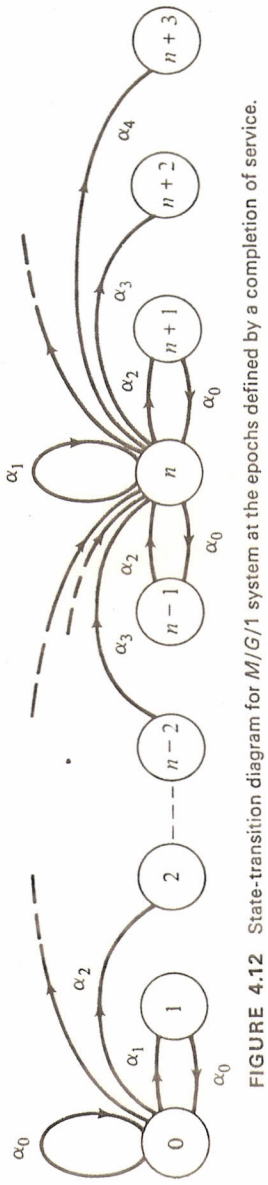

FIGURE 4.12 State-transition diagram for $M/G/1$ system at the epochs defined by a completion of service.

number of calls present at the time when a service is completed." Note that the state-transition diagram of Figure 4.12 is no longer of the birth-and-death type.

In the analysis that follows, we shall be interested only in the expected values $\bar{L}$, $\bar{W}$, $\bar{L}_q$, and $\bar{W}_q$, which can be derived without using the state-transition diagram. We shall, therefore, ignore Figure 4.12 from here on.

Let us define a random variable δ such that

$$\delta = \begin{cases} 0 & \text{if } N > 0 \\ 1 & \text{if } N = 0 \end{cases} \tag{4.66}$$

We can now write (4.62) in the form

$$N' = N + R - 1 + \delta \qquad \text{for all values of } N \geq 0 \tag{4.67}$$

Suppose now that the service time to the ith patient lasts exactly a time s. Then, from the properties of the Poisson process, it follows that

$$E[R \mid S = s] = \lambda s \tag{4.68}$$

$$E[R^2 \mid S = s] = \lambda^2 s^2 + \lambda s \tag{4.69}$$

It follows from (4.68) and (4.69) that the unconditional moments of r are

$$E[R] = \int_0^\infty E[R \mid S = s] f_S(s)\, ds = \int_0^\infty \lambda \cdot s f_S(s)\, ds = \lambda E[S] = \frac{\lambda}{\mu} \tag{4.70}$$

$$E[R^2] = \int_0^\infty E[R^2 \mid S = s] f_S(s)\, ds = \int_0^\infty (\lambda^2 s^2 + \lambda s) f_S(s)\, ds$$

$$= \lambda^2 E[S^2] + \lambda E[S] = \lambda^2 \left(\sigma_S^2 + \frac{1}{\mu^2}\right) + \frac{\lambda}{\mu} \tag{4.71}$$

Now, in the steady state we must have $E[N'] = E[N]$. But, from (4.67),

$$E[N'] = E[N] + E[R] - 1 + E[\delta]$$

or

$$E[\delta] = 1 - E[R] = 1 - \frac{\lambda}{\mu} = 1 - \rho \tag{4.72}$$

Note that (4.72) has a very real meaning: from the definition of δ, it follows that $E[\delta]$ is equal to the fraction of epochs, t_i, at which the queueing system will be found empty upon completion of a service. It follows that (4.72) cannot be meaningful unless[8] $\rho < 1$ ($\lambda < \mu$). This is also the condition under which steady state exists for $M/G/1$ systems.

[8] It can be shown that steady state is not reached when $\rho = 1$. Expected queue lengths and expected waiting times are infinite for that value of ρ.

Let us now square both sides of (4.67) to obtain

$$(N')^2 = N^2 + (R-1)^2 + \delta^2 + 2N(R-1) + 2N\delta + 2\delta(R-1)$$
$$= N^2 + (R-1)^2 + 2N(R-1) + \delta(2R-1) \tag{4.73}$$

where we have used the facts that $\delta^2 = \delta$ and that $2N\delta = 0$ (both resulting directly from the definition of δ). Taking now the expected values of both sides of (4.73) and noting that in the steady state $E[(N')^2] = E[N^2]$, we have

$$2E[N]E[1-R] = E[R^2] - 2E[R] + 1 + E[\delta]E[2R-1]$$

from which it follows, by also using (4.70)–(4.72), that

$$\bar{L}^* \triangleq E[N] = \rho + \frac{\rho^2 + \lambda^2\sigma_S^2}{2(1-\rho)} \tag{4.74}$$

We have used the asterisk in (4.74) to indicate that $\bar{L}^*$ denotes the expected number of users in the system *at the instants that follow the service completions* on which we have concentrated, the epochs.

It turns out that $\bar{L}^*$ is equal to $\bar{L}$, the expected number of calls in the queueing system that would be observed by someone arriving at a *random time* with the system in the steady state. To show this, it suffices to show that the steady-state pmf for the number in the system at the instants of service completion is identical to the steady-state pmf for the number in the system at any random instant. This we now proceed to do in a rather informal way. We define

π_n = steady-state probability that just after the completion of a service there are n calls left in the queueing system

P_n = steady-state probability that a call arriving at the system at a random time will find n calls in the queueing system

J_n = number of "downward jumps" from $n+1$ to n (in the number of calls in the queueing system) which are observed during a time interval T

$J = \sum_{n=0}^{\infty} J_n$ = total number of downward jumps in the number of calls in the queueing system observed during the time interval T

K_n = number of "upward jumps" from n to $n+1$ (in the number of calls in the queueing system) which are observed during the time interval T

$K = \sum_{n=0}^{\infty} K_n$ = total number of upward jumps in the number of calls in the queueing system observed during the time interval T

Obviously, downward jumps are due to service completions and upward jumps are due to call arrivals. Obviously, too, the quantities J_n and K_n can differ by at most 1 unit during any time interval T.

Assuming that the queueing system does reach steady state, we have, from the definition of π_n, that

$$\pi_n = \lim_{T\to\infty} \frac{J_n}{J} \tag{4.75}$$

Also, since steady state is reached, the number of upward jumps must be about the same as the number of downward jumps (i.e., the ratio of J to K must go to 1 as T goes to infinity). From this, plus the fact that J_n and K_n differ at most by one and from (4.75), we then have

$$\pi_n = \lim_{T\to\infty} \frac{J_n}{J} = \lim_{T\to\infty} \frac{K_n}{K} \tag{4.76}$$

The right-hand side above is the steady-state probability that the system is in state n at the instant of an arrival. Arrivals, however, occur in a Poisson manner, meaning that the instants of arrivals are completely random! Thus, $\lim_{T\to\infty}(K_n/K) = P_n$ and we have shown that

$$P_n = \pi_n \tag{4.77}$$

which is the desired result.

We can thus state now that

$$P_0 = \pi_0 = E[\delta] = 1 - \rho$$

and

$$\bar{L} = \sum_{n=0}^{\infty} n \cdot P_n = \sum_{n=0}^{\infty} n \cdot \pi_n = \bar{L}^*$$

and going back to (4.72) and (4.74), we have

$$P_0 = 1 - \rho \tag{4.78}$$

$$\bar{L} = \rho + \frac{\rho^2 + \lambda^2\sigma_S^2}{2(1-\rho)} \tag{4.79}$$

$$\bar{W} = \frac{\bar{L}}{\lambda} = \frac{1}{\mu} + \frac{\rho^2 + \lambda^2\sigma_S^2}{2\lambda(1-\rho)} \tag{4.80}$$

$$\bar{W}_q = \bar{W} - \frac{1}{\mu} = \frac{\rho^2 + \lambda^2\sigma_S^2}{2\lambda(1-\rho)} = \frac{\lambda[(1/\mu^2) + \sigma_S^2]}{2(1-\rho)} \tag{4.81}$$

$$\bar{L}_q = \lambda\bar{W}_q = \frac{\rho^2 + \lambda^2\sigma_S^2}{2(1-\rho)} \tag{4.82}$$

for the $M/G/1$ queueing system. These results are valid for steady-state conditions which exist whenever

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

Expressions (4.78)–(4.82) are remarkable for their simplicity, since they apply to *any* service-time distribution. [In fact, in deriving these expressions we made no assumptions whatsoever about the specific form of $f_S(s)$.] They are usually referred to as the "Pollaczek–Khintchine formulas." To use them, all that is needed to know about the service time is its expected value and its variance—which is certainly most convenient in practical applications. It is also important to note that $\bar{L}$ (as well as $\bar{W}$, $\bar{L}_q$, and $\bar{W}_q$) depends on the variance of the service times: increasing the consistency of service (i.e., reducing the variance of service times) improves the performance of the service facility.

Several additional results have been obtained with regard to the $M/G/1$ queueing system. For instance, from (4.78) we can conclude that the following ratio holds:

$$\frac{E[\text{length of busy period}]}{E[\text{length of idle period}]} = \frac{\text{fraction of time system is busy}}{\text{fraction of time system is idle}} = \frac{\rho}{1-\rho}$$

But since $E[\text{length of idle period}] = 1/\lambda$ (since we have Poisson arrivals with rate λ), it can be concluded that

$$E[\text{length of busy period}] = \frac{1}{\lambda}\frac{\rho}{1-\rho} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda} \qquad (4.83)$$

Interestingly, (4.83) is identical with (4.43), the expression for the expected length of the busy period for $M/M/1$ queueing systems.

For a given service time pdf, $f_S(s)$, it is also possible to obtain expressions (or numerical estimates) for the steady-state probabilities, P_n. This is accomplished by first obtaining an expression for the geometric transform of these probabilities (see, for instance, [GROS 74]).

Example 1 (continued)

To illustrate some of the above, let us take $X_0 = 2$ miles, $Y_0 = 1$ mile, $v_x = 30$ miles/hr, $v_y = 20$ miles/hr, $\bar{Z} = 10$ minutes, and $\sigma_Z^2 = 25$ minutes^2.

Solution

It follows from (4.60) and (4.61) that $1/\mu = 13.5$ minutes and $\sigma_S^2 = 27.1$ minutes^2. Thus, the service rate $\mu = 4.44$ calls/hr. We can then derive the following table, for different demand (call) rates, λ, under steady-state conditions:

λ (calls/hr)	ρ	P_0 (probab.)	$\bar{L}$ (no. calls)	$\bar{W}$ (min)	$\bar{L}_q$ (no. calls)	$\bar{W}_q$ (min)	$\bar{W}_{q,M}$ (min)	$\bar{W}_{q,D}$ (min)
0.5	0.1125	0.8875	0.1206	14.47	0.0081	0.97	1.71	0.86
1.0	0.225	0.775	0.2625	15.75	0.0375	2.25	3.92	1.96
1.5	0.3375	0.6625	0.4363	17.45	0.0988	3.95	6.88	3.44
2.0	0.45	0.55	0.6615	19.85	0.2115	6.35	11.05	5.52
2.5	0.5625	0.4375	0.9779	23.47	0.4154	9.97	17.36	8.68
3.0	0.675	0.325	1.4802	29.60	0.8052	16.10	28.04	14.02
3.5	0.7875	0.2125	2.4636	42.23	1.6761	28.73	50.03	25.01
4.0	0.9	0.1	5.5520	83.28	4.6520	69.78	121.50	60.75

The quantities P_0, $\bar{L}$, $\bar{W}$, $\bar{L}_q$, and $\bar{W}_q$ in this table have been computed by using relations (4.78)–(4.82). $\bar{W}_{q,M}$ and $\bar{W}_{q,D}$, the quantities listed in the two rightmost columns represent the average waiting time for the corresponding $M/M/1$ and $M/D/1$ systems, respectively. That is, $\bar{W}_{q,M}$ has been computed for a single-server system with negative exponential service time distribution and an expected service time, $1/\mu$, of 13.5 minutes; similarly, $\bar{W}_{q,D}$ corresponds to a constant service time equal to 13.5 minutes. Since an $M/M/1$ system can be viewed as just a special case of $M/G/1$, it is hardly surprising that when we set $\sigma_S^2 = 1/\mu^2$ (negative exponential service times) in (4.79)–(4.82), the expressions for the corresponding $M/M/1$ quantities are obtained. (Try one!) Similarly, the expressions for the $M/D/1$ system can be obtained by setting $\sigma_S^2 = 0$ in (4.79)–(4.82). A particularly simple and useful relationship to remember is that

$$\bar{W}_{q,D} = \frac{\bar{W}_{q,M}}{2} \qquad (4.84)$$

As one might expect from the fact that $\sigma_S^2 < 1/\mu^2$ in our example, the values of $\bar{W}_q$ for all values of λ in the table fall between the corresponding values of $\bar{W}_{q,M}$ and $\bar{W}_{q,D}$. In fact, there is a particularly convenient form for expressions (4.79)–(4.82) that brings out clearly the significance of the term that includes the variance of the service time: we can use the coefficient of variation $C_s(\triangleq \sigma_S/E[S] = \sigma_S \cdot \mu)$ for the service time to rewrite, say, (4.79) as

$$\bar{L} = \rho + \frac{\rho^2(1+C_s^2)}{2(1-\rho)} \qquad (4.79a)$$

(Remember that for negative exponential service times $C_s = 1$ and for constant service times $C_s = 0$.)

Finally, to conclude the example, we might want to review the table of numerical results to assess the performance of the ambulance service that we have described. It is interesting, for instance, that at an arrival rate of 1.5 calls/hr, the average delay before the ambulance is dispatched to a random call for service is about 4 minutes—longer than what it takes for the ambulance, on the average, to travel from the EMF to the point from which the call

has originated and back (= 3.5 minutes). And this, despite the fact that, for $\lambda = 1.5$, the ambulance is busy (traveling or at the scene of an incident) only about one third of the time. It is thus very likely that emergency medical service administrators would find the level of service (as manifested by the average dispatch delay, $\bar{W}_q$) provided by this single ambulance emergency medical system to be unacceptable for call rates greater than 1.5 or, at most, 2.0 calls/hr.

What to do, then? We might attempt to speed up service (reduce $1/\mu$) or "standardize" the service (reduce σ_S^2). Suppose that a 20 percent decrease could be achieved for $1/\mu$ [i.e., we could achieve $1/\mu' = (13.5)(0.8) = 10.8$ minutes]. Then for, say, $\lambda = 3.0$, we would obtain $W_q' = 7.82$ (assuming that σ_S^2 stays constant at 27.1). This is a better than 50 percent reduction in average dispatch delay!

Suppose, instead, that we could reduce the standard deviation of service times by 20 percent [i.e., that we could achieve $\sigma_S'' = (0.8)(27.1)^{1/2}$ or, in other words $(\sigma_S^2)'' = (0.64)(27.1) = 17.34$]. Then for $\lambda = 3.0$, and assuming that $1/\mu$ remains constant at 13.5, we would obtain $\bar{W}_q'' = 15.35$ minutes or an improvement of only about 5 percent over the original $\bar{W}_q$ of 16.1 minutes. In general, reductions in expected service times are usually much more effective than comparable reductions in the variance of service times. This should be obvious from the fact that changes in the expected service time, $1/\mu$ affect both the numerator and denominator of (4.79)–(4.82) by changing the utilization factor, ρ.

When it is not possible to reduce $1/\mu$ or σ_S^2 to achieve improved performance for a given demand rate λ, one has to resort to more drastic measures, such as increasing the number of ambulances in our present example or reducing the area of responsibility of the EMF (and thus λ as well). With m ambulances at hand ($m > 1$) that would mean an $M/G/m$ queueing system, which we proceed to discuss next. A more "complicated" spatially distributed $M/G/1$ queueing system will be discussed in Section 5.2.

4.8 USEFUL RESULTS FOR DIFFICULT-TO-ANALYZE QUEUEING SYSTEMS

4.8.1 Why Are $M/G/m$, $G/G/1$, and $G/G/m$ Difficult?

The natural extension of the $M/G/1$ model of Section 4.7 is the $M/G/m$ case, in which m independent and identical servers process users from a common queue with service times described by some "general" type of pdf. Unfortunately, unlike the transition from the $M/M/1$ to the $M/M/m$ case, which was straightforward, the transition from the $M/G/1$ to the $M/G/m$ model involves a "quantum jump" in the level of analytical complexity.

To see why, it is worthwhile to return to the $M/G/1$ model and review for a moment how our analysis of that queueing system proceeded. It will be

recalled that in that case we identified instants of time (the "epochs" $t_1, t_2, t_3, \ldots$, that coincide with the completion of service to users) such that, given the number, n, of users present at the queueing system at epoch t_k, one could immediately determine the probability that n' users will be present at epoch t_{k+1}. We did determine these state transition probabilities in (4.64) and (4.65) after deriving the expression for α_r, (4.63).

Let us now consider the $M/G/m$ case and attempt to write a relation such as (4.63) with the eventual aim of determining all state-transition probabilities for a state-transition diagram. Some thought will now convince the reader that the presence of m rather than 1 server makes it impossible to identify special instants of time (epochs) between which the $M/G/m$ system undergoes state transitions that can be described by easily obtainable state-transition probabilities. At the instants of the completion of a service, for instance, it does not suffice any more to know just the number of users in the system, as with the $M/G/1$ system. In order to specify the transition probabilities for the number of users at the next completion of a service, one also needs to know *how long each one* of the other servers who are occupied at the instant of the completion of a service has been busy with its occupant. This, of course, is not a concern in the $M/G/1$ case, where there are never any other servers which can be busy at the instant of a service completion.

A similar rationale applies to the case of the $G/G/1$ queuing system. Here there is only a single server, so, at the instant of a service completion, we do not have to worry about how long the other servers have been busy with their present occupants. However, interarrival gaps are now described by a general pdf with "memory." Therefore, at the time of a service completion, it is now necessary to know how long it has been since the last user arrival. That information is necessary to determine the probability that, say, r new users will arrive at the queueing system between the present service completion and the next one. Consequently, the transition probabilities, α_r, are also difficult to obtain for the $G/G/1$ system.

Finally, it follows a fortiori that $G/G/m$ systems pose an even worse problem to the analyst.

Contents of this section. We have indicated above that queueing systems of $M/G/m$, $G/G/1$, and $G/G/m$ types are difficult to analyze mathematically. As a consequence, the readily usable, general, and exact results about such systems are very few. On the other hand, several quite useful *approximate* results are available, and these will be the focus of our attention in the remainder of this section. It should be noted that most of these results have been developed in recent years, when much attention has been turned to the question of approximations in queueing theory.

In the following discussion we shall not derive or prove the validity of the expressions that will be presented. Appropriate references are provided for the interested reader.

4.8.2 *M/G/m* Queueing Systems with No Waiting Space

One of the few examples of "difficult" queueing systems for which exact results exist are *M/G/m* systems with no waiting space. Instances in which these models are, at least approximately, applicable in urban operations research are common and usually involve emergency services.

For example, a few years ago the following situation came under review in one of our major cities. A small fleet of hospital-affiliated ambulances were being used as the primary means of transporting emergency patients to local hospitals. At those times when *all ambulances were busy*, a back-up fleet of "ambulettes" run by the police department of that city was pressed into service. No queue was allowed to form for service by the hospital-affiliated ambulances. It was generally believed by EMS planners that police ambulettes provided inferior service (a belief that was apparently shared by many city residents). As a result, it was decided to increase the size of the hospital-affiliated fleet and it was desired to determine the minimum number of active hospital-affiliated ambulances that would be necessary to assure that:

$$P\{\text{a random emergency patient must be served by a police ambulette}\} \leq f$$

where f is a threshold probability ($0 < f < 1$).

Calls for emergency ambulance service arise in a Poisson fashion in time and the service times for calls (involving a trip to the location of the call, some time spent there, and, finally, transportation to a hospital) are random variables with "general" pdf's. Therefore, in the presence of a fleet of m hospital-affiliated ambulances, the fraction of calls serviced by police ambulettes is equal to the probability, P_m, that all m hospital ambulances are busy as given by a $M/G/m$ queueing model with no waiting space.

It has been shown (see, for instance, [GROS 74]) that, quite remarkably, the expressions for the steady-state probabilities for this $M/G/m$ case are *identical* to those for the corresponding $M/M/m$ system. Thus, as in (4.57), we have

$$P_n = \frac{(\lambda/\mu)^n/n!}{\sum_{i=0}^{m} (\lambda/\mu)^i/i!} \qquad 0 \leq n \leq m \tag{4.85}$$

and as a result the "loss formula" (the fraction of users who find the system full and are turned away) is, as before,

$$P_m = \frac{(\lambda/\mu)^m/m!}{\sum_{i=0}^{m} (\lambda/\mu)^i/i!} \tag{4.86}$$

To get back to our example, the number of hospital ambulances needed is the smallest m such that $P_m \leq f$. It is particularly surprising that (4.85) and (4.86) depend only on the mean of the service times.

M/G/∞ queueing system. The $M/G/\infty$ queueing system can now be viewed as a special case of the $M/G/m$ system with no waiting space, by taking the limiting value $m \to \infty$ for the latter system. Thus, we can use directly our last result. Replacing m by ∞ in (4.85), we have

$$P_n = \frac{(\lambda/\mu)^n e^{-\lambda/\mu}}{n!} \qquad 0 \leq n \tag{4.87}$$

As one would expect [since (4.85) is identical to (4.57)], this expression is the same as (4.50) for the $M/M/\infty$ system. As with the $M/M/\infty$ system, we also have $\bar{L} = \lambda/\mu$, $\bar{W} = 1/\mu$, and $\bar{L}_q = \bar{W}_q = 0$ in this case.

It is instructive to derive (4.87) directly because in the process one can also derive the time-dependent probabilities $P_n(t)$ of having n users in the $M/G/\infty$ system at time t assuming that the system was empty at $t = 0$. In the following we shall use $F_S(s)$ to denote the cdf for the service times and also use the fact that

$$E[S] = \mu^{-1} = \int_0^\infty [1 - F_S(s)]\, ds$$

Exercise 4.6 Show that the above relationship holds for any random variable S that assumes only non-negative values. [Hint: It is easier to work initially with a discrete random variable.]

We define

1. $P_n(t) \equiv P\{\text{at time } t \text{ there are exactly } n \text{ users being serviced}\}$.
2. $P_n \equiv \lim_{t\to\infty} P_n(t)$.

We wish to prove the following:

$$P_n(t) = \frac{\left\{\lambda \int_0^t [1 - F_S(x)]\, dx\right\}^n \exp\left[-\lambda \int_0^t (1 - F_S(x))\, dx\right]}{n!}, \qquad n = 0, 1, 2, \ldots \tag{4.88}$$

$$P_n = \frac{(\lambda/\mu)^n e^{-(\lambda/\mu)}}{n!}, \qquad n = 0, 1, 2 \tag{4.87}$$

[Note from (4.88) that, for any given t, $P_n(t)$ is a Poisson pmf.]