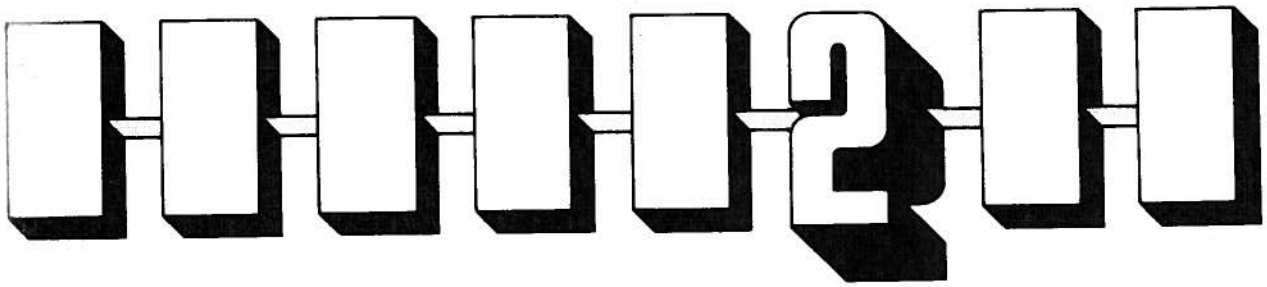

Introduction to Queueing Theory

2–1 Poisson Process
2–2 The M/M/1 Queue
2–3 Little's Formula, $L = \lambda W$
2–4 State-dependent Queues: Birth-death Processes
2–5 M/G/1 Queue: Mean Value Analysis
2–6 Nonpreemptive Priority Queueing Systems

This book focuses on the *performance* analysis of data networks. Although a great deal of qualitative material, describing real networks and network architectures, appears, the emphasis where possible is on quantitative considerations. These considerations involve the interplay among various performance parameters and how they relate to network resources that are to be controlled.

As noted in Chapter 1, two generic types of networks are considered: packet-switched and circuit-switched networks. In the packet-switched case, packets—blocks of data of varying length—are transmitted over a network from source to destination, following some routing path prescribed as part of the network design. The transmission facilities are shared by packets as they traverse the network. In the circuit-switched case, a transmission path end to end is set up for a pair of users who desire to establish a call. (The data flowing could equally well be voice or data messages.) The number and length of packets entering or traversing a network at any time, the number of calls arriving at a network entrance point in a given time, the length (the holding time of these calls)—all of these parameters generally vary statistically. In order to come up with quantitative measures of performance, therefore, probabilistic concepts must be used to study their interaction with a network. Queueing theory plays a key role in the analysis of networks, and this chapter covers the basic princi-

ples of queueing in order to prepare the reader for the quantitative material that follows.

Queueing arises very naturally in a study of packet-switched networks: Packets, arriving either at an entry point to the network or at an intermediate node on the way to the destination, are buffered, processed to determine the appropriate outgoing transmission link connected to the next node along the path, and then read out over that link when their time for transmission comes up. The time spent in the buffer waiting for transmission is a significant measure of performance for such a network, since end-to-end time delay, of which this wait time is a component, is one element of the network performance experienced by a user. The wait time depends, of course, on nodal processing time and packet length. It also depends on the transmission link capacity, in packets/sec capable of being transmitted, on the traffic rate in packets/sec arriving at the node in question, and on the service discipline used to handle the packets. Our queueing theory formulation will in fact consider most of these items. In our quantitative study of packet-switched networks we shall neglect nodal processing time for the most part, for the sake of simplicity. In our study of circuit-switched networks in Chapter 11, however, we shall consider the processing of calls at a node.

Queueing theory also arises naturally in the study of circuit-switched networks, not only in studying the processing of calls but in analyzing the relation at a node or switching exchange between trunks available (each capable of handling one call) and the probability that a call to be set up will be blocked or queued for service at a later time. In fact, historically, much of modern queueing theory developed out of telephone traffic studies at the beginning of the twentieth century. Integrated networks, which combine aspects of packet switching and circuit switching, will be considered later in this book, and the discussion there will of necessity involve the use of queueing concepts.

Consider the simplest model of a queue, as depicted in Fig. 2–1. To keep the discussion concrete, the queue in this case is shown handling packets of data. These packets could also be calls queueing up for service in a circuit-switched system. More generally, in the queueing literature jargon, they would be "cus-

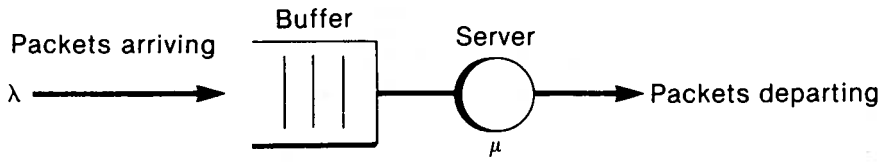

Figure 2–1 Model of a single-server queue

tomers" queueing up for service. The packets arrive randomly, at an average rate of λ packets/time (we shall use units of packets/sec most often). They queue up for service in the buffer shown and are then served, following some specified service discipline, at an average rate of μ packets/time. In the example of Fig. 2-1, only a single server is shown. More generally, multiple servers may be available, in which case more than one packet may be in service at any one time.

The concept of a server is of course well known to all of us from innumerable waits at the supermarket, bank, movie house, automobile toll booth, and so forth. In the context of a data network, the server is the transmission facility — an outgoing link, line, or trunk (all three terms will be used interchangeably in the material following) that transmits data at a prescribed digital rate of C data units/time. Most frequently, the data units are given in terms of bits or characters, and one talks of a transmission rate or link capacity C in units of bits/sec or characters/sec. A transmission link handling 1000-bit packets and transmitting at a rate of $C = 2400$ bps, for example, would be capable of transmitting at a rate of $\mu = 2.4$ packets/sec. More generally, if the average packet length in bits is $1/\mu'$ bits, and is given in units of bits/packet, $\mu = \mu' C$ packets/sec is the transmission capacity in units of packets/sec. (For circuit-switched calls, the "customer" would be a call; λ arrivals/sec represents the average *call* arrival rate, or the number of calls/sec handled on the average. The parameter $1/\mu$, in units of sec/call, is called the average call holding time.)

It is apparent that as the packet arrival rate λ approaches the packet rate capacity μ, the queue will start to build up. For a finite buffer (the situation in real life), the queue will eventually saturate at its maximum value as λ exceeds μ and continues to increase. If the buffer is filled, all further packets (customers) are blocked on arrival. We shall demonstrate this phenomenon quantitatively later in this chapter. If for simplicity the buffer is assumed to be infinite (an assumption we shall make often to simplify analysis), the queue becomes unstable as $\lambda \rightarrow \mu$. We will show that $\lambda < \mu$ to ensure stability in this case of a single-server queue. In particular, we shall find the parameter $\rho \equiv \lambda/\mu$ playing a critical role in queueing analysis. This parameter is often called the link *utilization* or *traffic intensity*. Note that it is defined as the ratio of load on the system to capacity of the system. For a single-server queue, as ρ approaches and exceeds one, the region of congestion is encountered, time delays begin to increase rapidly, and packets arriving are blocked more often.

To quantify the discussion of time delay, blocking performance, and packet throughput (the actual number of packets/time that get through the system), and their connection with both μ (the packet rate capacity) and the size of the buffer in Fig. 2-1, one needs a more detailed model of the queueing system. Specifically, those performance parameters among others will be shown to depend on the probabilities of state of the queue. The state is in turn defined to be the number of packets on the queue (including the one in service if the queue is

nonempty). To calculate the probabilities of state, one must have knowledge of

1. The packet arrival process (the arrival statistics of the incoming packets),
2. The packet length distribution (this is called the service-time distribution in queueing theory), and
3. The service discipline (for example, FCFS—first come–first served—or FIFO service; LCFS—last come–first served; or some form of priority discipline).

For multiple-server queues, the state probabilities depend on the number of servers as well. (The servers represent the trunks or outgoing links simultaneously transmitting packets, or handling calls in the case of a circuit-switched system.)

In most of our work in this book we model the packet- and call-arrival processes as *Poisson processes.* The Poisson process is the most frequently used arrival process in queueing theory. For this reason we devote the next section to a brief discussion of this process and show how it is related intimately to exponential statistics. The simplest queueing system to analyze, the so-called M/M/1 queue, is one with Poisson arrivals and an exponential service-time distribution. It is easy to obtain the probabilities of state of this queueing system for both the finite and the infinite queue cases, as shown in Section 2–2. We then derive a simple but general relation between average time delay and average number of customers (packets or calls) in a queue, called Little's formula (Section 2–3). This relation will be found useful throughout the remainder of this book.

Continuing our introduction to queueing theory in this chapter, we then present two sections that generalize the M/M/1 queue analysis. In Section 2–4 we show how one can analyze state-dependent queues. (This material will be found particularly useful in the discussion of circuit switching in the latter part of the book.) In Section 2–5 we consider a queue with a general service time distribution and Poisson arrivals, the so-called M/G/1 queue. This enables us, as a special case, to determine the effect of fixed length packets. More generally, we derive a very interesting and extremely useful expression for the average time delay of a queue with general service (packet length or call-holding time) statistics. The result appears as a simple modification of the M/M/1 (exponential service time) time delay result. A brief discussion of priority queueing for the single-server case follows the M/G/1 analysis.

2–1 Poisson Process

As noted in the previous section, the Poisson process is the arrival process used most frequently to model the behavior of queues. It has been used extensively in

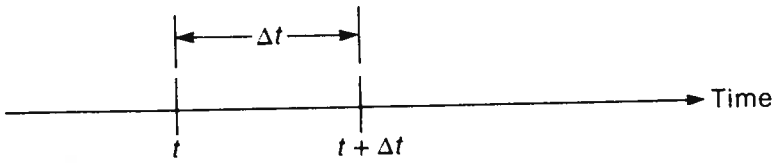

Figure 2-2 Time interval used in defining Poisson process

telephone traffic as well as in evaluating the performance of telephone switching systems and computer networks. Some of these analyses will be presented later in this book. In addition, the Poisson process has been used to model both photon generation and photodetector statistics, to represent shot noise processes, and to study semiconductor electron-hole generation phenomena, among other applications.

Three basic statements are used to define the Poisson arrival process. Consider a small time interval Δt ($\Delta t \to 0$), separating times t and $t + \Delta t$, as shown in Fig. 2-2. Then,

1. The probability of one arrival in the interval Δt is defined to be $\lambda \Delta t + o(\Delta t)$,* $\lambda \Delta t \ll 1$, and λ a specified proportionality constant.
2. The probability of zero arrivals in Δt is $1 - \lambda \Delta t + o(\Delta t)$.
3. Arrivals are memoryless: An arrival (event) in one time interval of length Δt is independent of events in previous or future intervals.

With this last definition the Poisson process is seen to be a special case of a *Markov* process, one in which the probability of an event at time $t + \Delta t$ depends on the probability at time t only [PAPO], [COX]. Note that with defining relations 1 and 2, more than one arrival or occurrence of an event in the interval Δt ($\Delta t \to 0$) is ruled out, at least to $0(\Delta t)$.

If one now takes a larger finite interval T, one finds the probability $p(k)$ of k arrivals in T to be given by

$$p(k) = (\lambda T)^k e^{-\lambda T}/k! \qquad k = 0, 1, 2, \ldots \tag{2-1}$$

* $o(\Delta t)$ implies that other terms are higher order in Δt and that they go to zero more rapidly than Δt as $\Delta t \to 0$.

[PAPO] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 2d ed., 1984.

[COX] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*, Methuen, London, 1965.

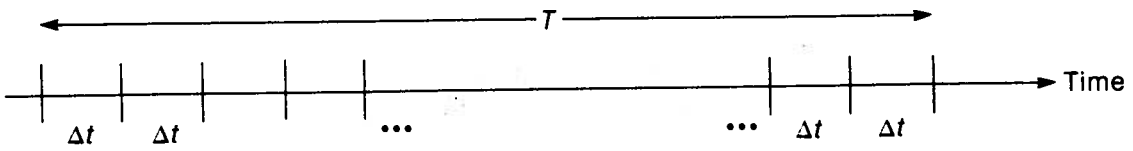

Figure 2-3 Derivation of Poisson distribution

This is called the *Poisson distribution.* It is left to the reader to show that this distribution is properly normalized ($\sum_{k=0}^{\infty} p(k) = 1$) and that the mean or expected value is given by

$$E(k) = \sum_{k=0}^{\infty} kp(k) = \lambda T \tag{2-2}$$

The variance $\sigma_k^2 \equiv E[k - E(k)]^2 = E(k^2) - E^2(k)$ turns out to be given by

$$\sigma_k^2 = E(k) = \lambda T \tag{2-3}$$

This is also left as an exercise to the reader. The parameter λ, defined originally as a proportionality constant (see defining relation 1 for the Poisson process), turns out to be a rate parameter:

$$\lambda = E(k)/T$$

from Eq. (2-2). It thus represents the average rate of Poisson arrivals, as implied in our discussion in the previous section (see Fig. 2-1).

From Eqs. (2-2) and (2-3) it is apparent that the standard deviation σ_k of the distribution, normalized to the average value $E(k)$, tends to zero as λT increases: $\sigma_k/E(k) = 1/\sqrt{\lambda T}$. This implies that for large λT the distribution is closely packed about the average value λT. If one thus actually measures the (random) number of arrivals n in a large interval T ("large" implies $\lambda T \gg 1$, or $T \gg 1/\lambda$), n/T should be a good estimate of λ. Note also that $p(0) = e^{-\lambda T}$. As λT increases, with the distribution peaking eventually about $E(k) = \lambda T$, the probability of *no* arrivals in the interval T approaches zero exponentially with T.

The Poisson distribution of Eq. (2-1) is easily derived using the three defining relations of the Poisson process. Referring to Fig. 2-3, consider a sequence of m small intervals, each Δt units long. Let the probability of one event (arrival) in any interval Δt be $p = \lambda \Delta t$, while the probability of 0 events is $q = 1 - \lambda \Delta t$. Using the memoryless (independent) relation, it is then apparent that the probability of k events (arrivals) in the interval $T = m\Delta t$ is given by the *binomial* distribution

$$p(k) = \binom{m}{k} p^k q^{m-k} \tag{2-4}$$

with

$$\binom{m}{k} \equiv m!/(m-k)!k!$$

Now let $\Delta t \to 0$, but with $T = m\Delta t$ fixed. Using the defining equation for the exponential,

$$\lim_{t \to 0} (1 + at)^{k/t} = e^{ak}$$

and approximating factorial terms by the Stirling approximation, one gets Eq. (2-1). Details are left to the reader.

Now consider a large time interval, and mark off the times at which a Poisson event (arrival) occurs. One gets a random sequence of points as shown in Fig. 2-4. The time between successive arrivals is represented by the symbol τ. It is apparent that τ is a continuously distributed positive random variable. For Poisson statistics, it turns out that τ is an *exponentially distributed* random variable; i.e., its probability density function $f_\tau(\tau)$ is given by

$$f_\tau(\tau) = \lambda e^{-\lambda\tau} \qquad \tau \geq 0 \tag{2-5}$$

This exponential interarrival distribution is sketched in Fig. 2-5. For Poisson arrivals the time between arrrivals is thus more likely to be small than large, the probability between two successive events decreasing exponentially with the time τ between them.

A simple calculation indicates that the mean value $E(\tau)$ of this exponential distribution is

$$E(\tau) = \int_0^\infty \tau f_\tau(\tau) d\tau = 1/\lambda \tag{2-6}$$

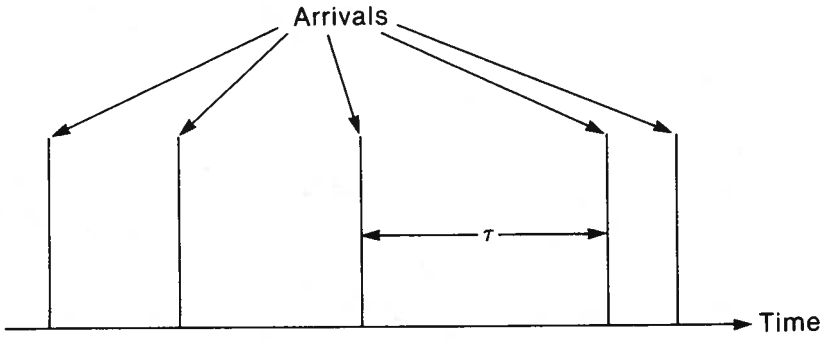

Figure 2-4 Poisson arrivals

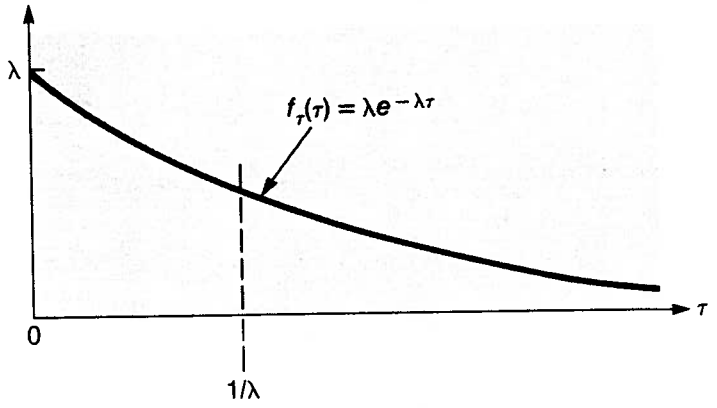

Figure 2–5 Exponential interarrival distribution

while its variance is given by

$$\sigma_\tau^2 = 1/\lambda^2 \tag{2-7}$$

The average time between arrivals is as expected, for if the rate of arrivals is λ, the time between arrivals should be $1/\lambda$.

The fact that Poisson arrival statistics give rise to an exponential interarrival distribution is easily deduced from the Poisson distribution of Eq. (2–1).

Consider the time diagram of Fig. 2–6. Let τ be the random variable representing the time to the first arrival after some arbitrary time origin, as shown. Take any value x. *No* arrivals occur in the interval $(0, x)$ if and only if $\tau > x$. The probability that $\tau > x$ is just the probability that no arrivals occur in

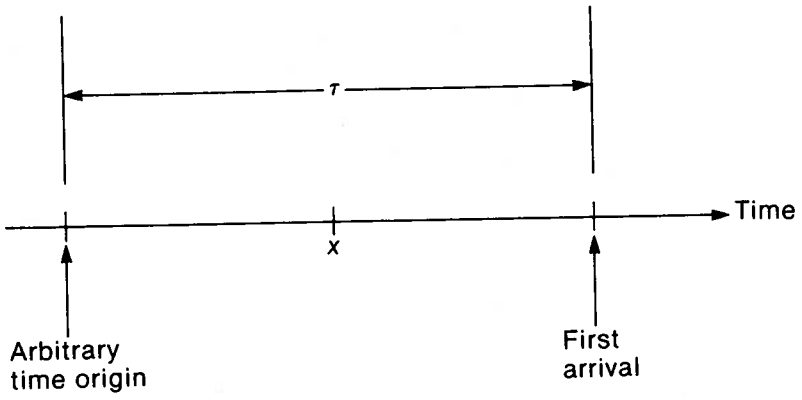

Figure 2–6 Derivation of exponential distribution

$(0, x)$; i.e.,

$$P(\tau > x) = \text{prob. (number of arrivals in } (0, x) = 0)$$
$$= e^{-\lambda x}$$

from Eq. (2-1). Then the probability that $\tau \leq x$ is

$$P(\tau \leq x) = 1 - e^{-\lambda x}$$

But this is just the cumulative probability distribution $F_\tau(x)$ of the random variable τ. Hence we have

$$F_\tau(x) = 1 - e^{-\lambda x} \tag{2-8}$$

from which the probability density function $f_\tau(x) = dF_\tau(x)/dx = \lambda e^{-\lambda x}$ follows immediately.

The close connection between the Poisson arrival process and the exponential interarrival time can be exploited immediately in discussing properties of the *exponential service-time distribution.* Thus consider a queue with a number of customers (packets or calls) waiting for service. Focus attention on the output of the queue, and mark the time at which a customer completes service. This is shown schematically in Fig. 2-7. Let the random variable representing time between completions be r, as shown. This must also be the service time if the next customer is served as soon as the one in service leaves the system. In particular, take the case where r is exponentially distributed in time, with an average value $E(r) = 1/\mu$. Thus

$$f_r(r) = \mu e^{-\mu r} \qquad r \geq 0 \tag{2-9}$$

But comparing Fig. 2-7 with Fig. 2-4, it is apparent that if r, the time between completions, is exponential, the completion times themselves must represent a Poisson process! The service process is the complete analog of the arrival process. On this basis, the probability of a service completion in the small time

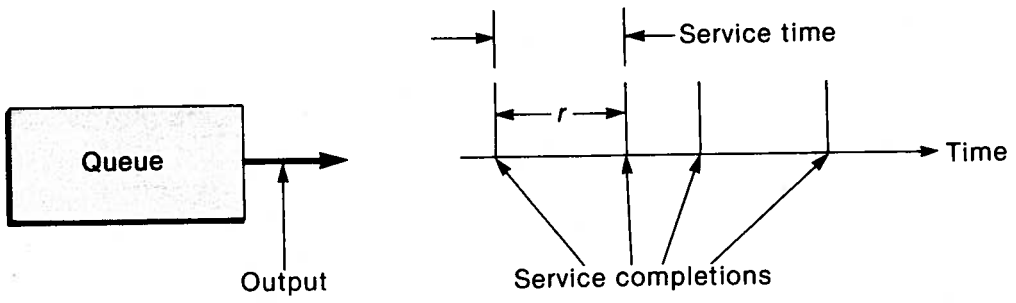

Figure 2-7 Service completions at output of a queue

interval $(t, t + \Delta t)$ is just $\mu\Delta t + 0(\Delta t)$, whereas the probability of *no* completion in $(t, t + \Delta t)$ is $1 - \mu\Delta t + 0(\Delta t)$, independent of past or future completions. The exponential model for service carries with it the memoryless property used as one of the defining relations of the Poisson process.

Before going on to our study of queueing, we introduce an additional property of the Poisson process. Say that m independent Poisson streams, of arbitrary rates $\lambda_1, \lambda_2, \ldots, \lambda_m$, respectively, are merged. Then the composite stream is itself Poisson, with rate parameter $\lambda = \sum_{i=1}^{n} \lambda_i$. This is an extremely useful property, and is one of the reasons that Poisson arrivals are often used to model arrival processes. In the context of packet-switched and circuit-switched networks, this situation occurs when combining statistically packets or calls from a number of data sources (terminals or telephones), each of which generates packets or calls (as the case may be) at a Poisson rate. A simple proof is as follows: Let $N^{(i)}(t, t + \Delta t)$ be the number of events in Poisson process i, $i = 1, 2, \ldots, m$ in the interval $(t, t + \Delta t)$. Let $N(t, t + \Delta t)$ be the total number of events from the composite stream. Then

$$\begin{aligned} \text{prob. } [N(t, t + \Delta t) = 0] &= \prod_{i=1}^{n} \text{prob. } [N^{(i)}(t, t + \Delta t) = 0] \\ &= \prod_{i=1}^{m} [1 - \lambda_i \Delta t + 0(\Delta t)] = 1 - \lambda\Delta t + 0(\Delta t), \end{aligned} \tag{2-10}$$

$\lambda = \sum_{i=1}^{m} \lambda_i$, since the individual processes are independent. A similar calculation shows that

$$\text{prob. } [N(t, t + \Delta t) = 1] = \lambda\Delta t + 0(\Delta t) \tag{2-11}$$

This proves the desired relation.

Sums of Poisson processes are thus distribution conserving: They retain the Poisson property. This property will be used implicitly in a number of places in this book in working out examples of queueing and buffering calculations.

2-2 The M/M/1 Queue

We now use the material of the previous section, on the Poisson process, to determine the properties of the simplest model of a queue, the M/M/1 queue. This is a queue of the single-server type, with Poisson arrivals, exponential service-time statistics, and FIFO service. The notation M/M/1 used is due to British statistician D. G. Kendall. The Kendall notation for a general queueing system is of the form A/B/C. The symbol A represents the arrival distribution, B represents the service distribution, and C denotes the number of servers used.

The symbol M in particular, from the Markov process, is used to denote the Poisson process or the equivalent exponential distribution. An M/M/m queue is thus one with Poisson arrivals, exponential service statistics, and m servers. An M/G/1 queue has Poisson arrivals, *general* service distribution, and a single server. A special case is the M/D/1 queue, with D used to represent fixed (deterministic) or *constant* service time. We shall have more to say about these other queueing structures in later sections.

As noted in the introduction to this chapter, the statistical properties of the M/M/1 queue, the average queue occupancy, the probability of blocking for a finite queue, average throughput, and so forth, are readily determined once we find the probabilities of state p_n at the queue. By definition, p_n is the probability that there are n customers (packets or calls) in the queue, including the one in service. By implication the system is operating at steady state so that these probabilities do not vary with time. Starting from some initial defined values (for example, an empty queue state), one expects these probabilities to approach steady-state, stationary, non-time-varying values as time goes on, if the arrival and service-time distributions are invariant with time. We shall show later on that these stationary probabilities are readily determined from simple flow balance arguments. At this point we use a more general, time-dependent argument.

Specifically, let the arrival process to the single-server queue of Fig. 2-8 be Poisson, with parameter λ. Let the service-time process (packet length or call holding time) be exponential, with parameter μ, as shown. Then the probability $p_n(t + \Delta t)$ that there are n customers (packets or calls) in the queue at time $(t + \Delta t)$ may readily be found in terms of corresponding probabilities at time t. Referring to the state-time diagram of Fig. 2-9, it is apparent that if the queue is in state n at time $t + \Delta t$, it could only have been in states $n - 1$, n, or $n + 1$ at time

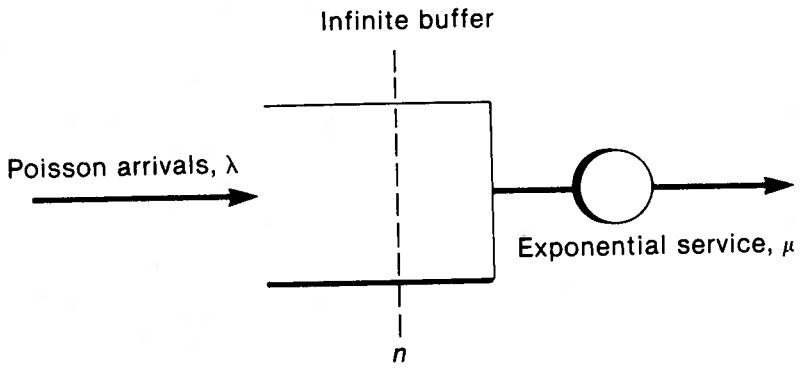

Figure 2-8 M/M/1 queue

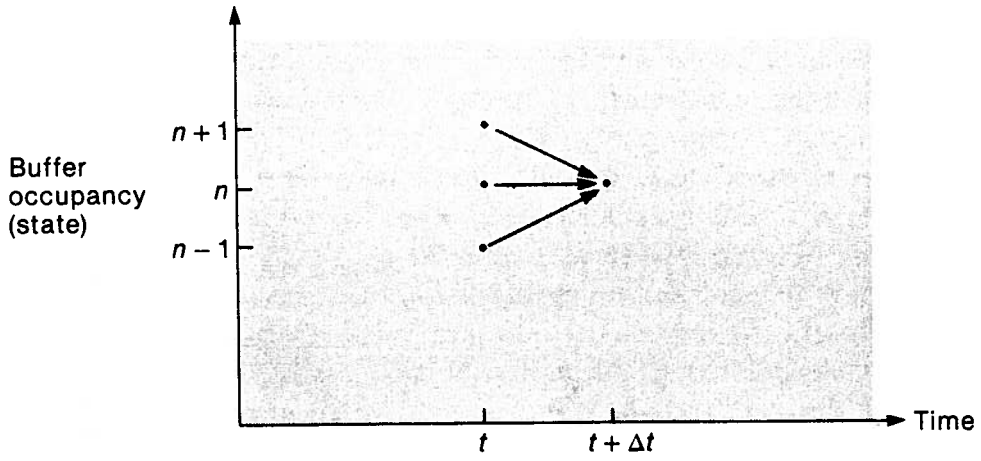

Figure 2-9 M/M/1 state-time diagram

t. (We assume here for simplicity that $n \geq 1$.) The probability $p_n(t + \Delta t)$ that the queue is in state n at time $t + \Delta t$ must be the sum of the (mutually exclusive) probabilities that the queue was in states $n - 1$, n, or $n + 1$ at time t, each multiplied by the (independent) probability of arriving at state n in the intervening Δt units of time. We thus have, as the generating equation for $p_n(t + \Delta t)$,

$$\begin{aligned} p_n(t + \Delta t) = {} & p_n(t)[(1 - \lambda\Delta t)(1 - \mu\Delta t) + \mu\Delta t \cdot \lambda\Delta t + 0(\Delta t)] \\ & + p_{n-1}(t)[\lambda\Delta t(1 - \mu\Delta t) + 0(\Delta t)] \\ & + p_{n+1}(t)[\mu\Delta t(1 - \lambda\Delta t) + 0(\Delta t)] \end{aligned} \quad (2-12)$$

The transition probabilities of moving from one state to another have been obtained by considering the ways in which one could move between the two states and calculating the respective probabilities, using the properties of the arrival and service-time distributions. As an example, if the system remains in state n, $n \geq 1$, there could have been either one departure and one arrival, with probability $\mu\Delta t \cdot \lambda\Delta t$, or no departure and no arrival, with probability $(1 - \mu\Delta t)(1 - \lambda\Delta t)$, as shown. The other terms in Eq. (2-12) are obtained similarly.

Since $0(\Delta t)$ includes terms of order $(\Delta t)^2$ and higher, the terms involving $(\Delta t)^2$ in Eq. (2-12) should be incorporated in $0(\Delta t)$. (They were retained and shown explicitly to help the reader understand Eq. (2-12).) Simplifying Eq. (2-12) in this way, and dropping $0(\Delta t)$ terms altogether, one gets

$$p_n(t + \Delta t) = [1 - (\lambda + \mu)\Delta t]p_n(t) + \lambda\Delta t\, p_{n-1}(t) + \mu\Delta t p_{n+1}(t) \quad (2-12a)$$

Eq. (2-12a) can be used to study the time-dependent (transient) behavior of the M/M/1 queue given that the queue is started at time $t = 0$ in some known state

or set of states. Alternatively, a differential-difference equation governing the time variation of $p_n(t)$ may be found by expanding $p_n(t + \Delta t)$ in a Taylor series about t and retaining the first two terms only:

$$p_n(t + \Delta t) \doteq p_n(t) + \frac{dp_n(t)}{dt}\Delta t \tag{2-13}$$

Using Eq. (2-13) in Eq. (2-12a) and simplifying, one readily obtains the following equation:

$$\frac{dp_n(t)}{dt} = -(\lambda + \mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t) \tag{2-14}$$

This is the differential-difference equation to be solved to find the time variation of $p_n(t)$ explicitly.

As stated earlier, $p_n(t)$ should approach a constant, stationary value p_n as time goes on. Assuming that this is the case (it will be shown later that the condition $\lambda < \mu$ ensures this in the case of the infinite queue), we must have $dp_n(t)/dt = 0$ at the stationary value of p_n as well. Equation (2-14), for the case of stationary, non-time-varying probabilities, then simplifies to the following equation involving the stationary state probabilities p_n of the M/M/1 queue:

$$(\lambda + \mu)p_n = \lambda p_{n-1} + \mu p_{n+1} \qquad n \geq 1 \tag{2-15}$$

This equation can be given a physical interpretation that enables us to write it down directly, by inspection, without going through the lengthy process of deriving it from Eq. (2-12). More important, using the approach to be described, we shall be able to write similar equations down, by inspection, for more general state-dependent queues later in this chapter. More complex queueing systems arising in the study of both packet- and circuit-switched networks will be analyzed in a similar manner in later chapters.

Consider the state diagram of Fig. 2-10, which represents the M/M/1 queue. Because of the Poisson arrival and departure processes assumed, transitions between adjacent states only can take place with the rates shown. There is a rate λ of moving up one state due to arrivals in the system, whereas there is a rate μ of moving down due to service completions or departures. Alternatively, if one multiples the rates by Δt, one has the probability $\lambda\Delta t$ of moving up one state due to an arrival, or the probability $\mu\Delta t$ of dropping down one state due to a service completion (departure). (If the system is in state 0, i.e., it is empty, it can only move up to state 1 due to an arrival.)

The form of Eq. (2-15) indicates that there is a stationary balance principle at work: The left-hand side of Eq. (2-15) represents the rate of *leaving* state n, given the system was in state n with probability p_n. The right-hand side represents the rate of *entering* state n, from either state $n - 1$ or state $n + 1$. In order for stationary state probabilities to exist, the two rates must be equal.

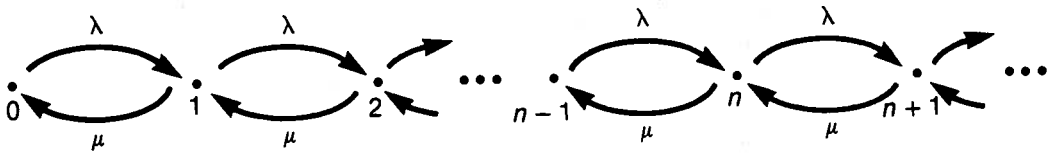

Figure 2-10 State diagram, M/M/1 queue

Balance equations play a key role in the study of queueing systems. We shall encounter similar equations later in this chapter in studying state-dependent queues; in Chapter 5, in studying queueing networks; and in later chapters, in studying more complex queueing systems arising as models of data networks. The relation of balance equations to time-invariant or equilibrium state probabilities is explored rigorously and at length in the seminal book by Kelly [KELL].

The solution of Eq. (2-15) for the state probabilities can be carried out in a number of ways. The simplest way is to again apply balance arguments. Consider Fig. 2-11, which represents the state diagram of the M/M/1 queue again drawn as in Fig. 2-10, but with two closed "surfaces," 1 and 2, sketched as shown. If one calculates the total "probability flux" crossing surface 1, and equates the flux leaving (rate of leaving state n) to the flux entering (rate of entering state n), one gets Eq. (2-15). Now focus on surface 2, which encloses the entire set of points from 0 to n. The flux entering the surface is μp_{n+1}; the flux leaving is λp_n. Equating these two, one gets

$$\lambda p_n = \mu p_{n+1} \tag{2-16}$$

It is left to the reader to show that this simple balance equation does in fact satisfy Eq. (2-15). The intuitive concept of balancing rates of departure from a state to rates of entering that state thus not only allows a balance equation to be set up (Eq. (2-15)), but a solution to be obtained as well! Repeating Eq. (2-16) n times, one finds very simply that

$$p_n = \rho^n p_0 \qquad \rho \equiv \lambda/\mu \tag{2-17}$$

To find the remaining unknown probability p_0, one must now invoke the probability normalization condition

$$\sum_n p_n = 1$$

[KELL] F. P. Kelly, *Reversibility and Stochastic Networks*, John Wiley & Sons, Chichester, U.K., 1979.

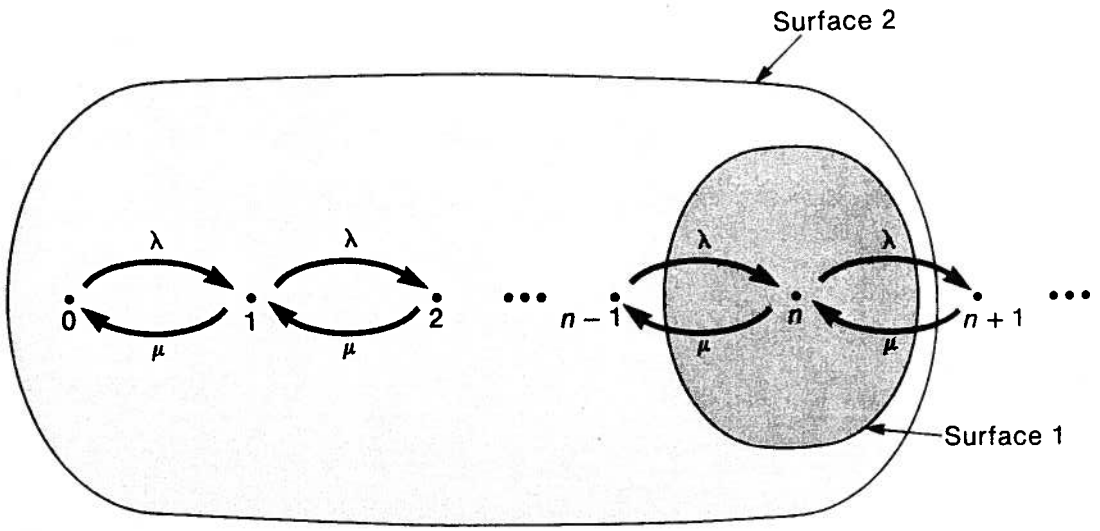

Figure 2-11 Flow balance, M/M/1 queue

For the case of an infinite M/M/1 queue, one finds very simply that $p_0 = (1 - \rho)$, *if* $\rho < 1$, and one gets, finally,

$$p_n = (1 - \rho)\rho^n \qquad \rho \equiv \lambda/\mu < 1 \tag{2-18}$$

as the equilibrium state probability solution for the M/M/1 queue. Note the necessary condition $\rho = \lambda/\mu < 1$, which was alluded to earlier. Its significance, again, is that for equilibrium to exist the arrival rate or load on the queue must be less than the *capacity* μ. If this condition is violated for this infinite queue model, the queue continues to build up in time, and equilibrium is never reached. Mathematically, the fact that equilibrium exists only for $\rho < 1$ is noted by considering the limiting case $\rho = 1$. Since $p_0 = (1 - \rho)$, $p_0 = 0$. But from Eq. (2-16), $p_1 = \rho p_0 = 0$, $p_2 = 0$, ... All the stationary probabilities are thus zero, a contradiction, and equilibrium does not exist. More detailed and general discussions of conditions for equilibrium in the context of Markov processes appear in [COX] and [KELL].

The M/M/1 state probability distribution of Eq. (2-18) is called a *geometric distribution*. An example, for $\rho = 0.5$, appears in Fig. 2-12. Note that since the probability the queue is empty is $p_0 = 1 - \rho$, the probability that the queue is nonempty is just ρ, the utilization.

Now consider the extension of this analysis to the case of a *finite queue*, accommodating at most N packets. A little thought indicates that the governing balance equation, Eq. (2-15), is unchanged except for the two boundary points

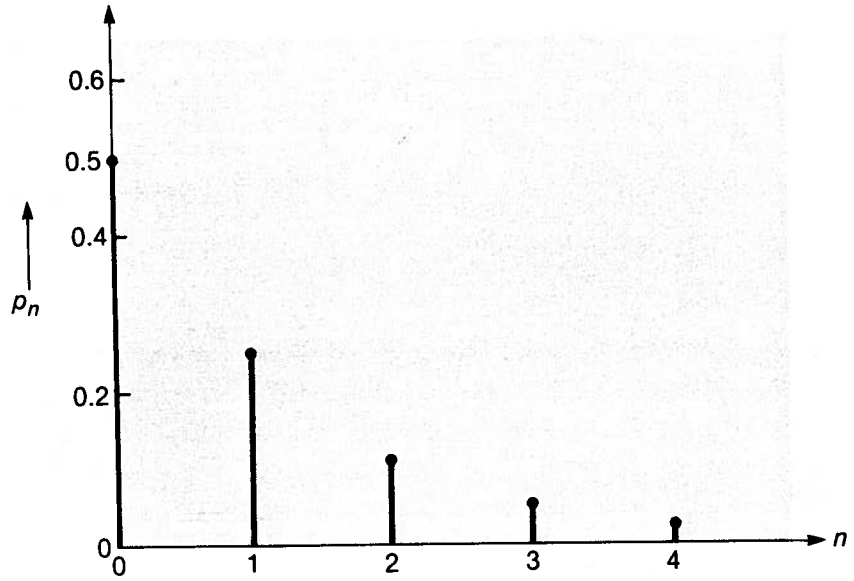

Figure 2-12 M/M/1 state probabilities, $\rho = 0.5$

$n = 0$ and $n = N$. Equation (2-17) is still the solution. The only difference now is that the probability p_0 that the queue is empty changes to accommodate the normalization condition summed over a finite number of states:

$$\sum_{n=0}^{N} p_n = 1$$

It is left for the reader to show that

$$p_0 = (1 - \rho)/(1 - \rho^{N+1}) \tag{2-19}$$

in this case, so that

$$p_n = (1 - \rho)\rho^n/(1 - \rho^{N+1}) \tag{2-20}$$

for the finite M/M/1 queue.

In particular, the probability that the queue is full is p_N, given by

$$p_N = (1 - \rho)\rho^N/(1 - \rho^{N+1}) \tag{2-21}$$

But this should be the same as the probability of blocking: the probability that customers (packets or calls) are turned away and not accepted by the queue. This may be demonstrated by the following simple argument, which will be found useful later in some of our data network performance analysis. Consider the picture of a queue shown in Fig. 2-13. This does not have to be a finite M/M/1 queue. It can be *any* queueing system that blocks customers on arrival. A load λ,

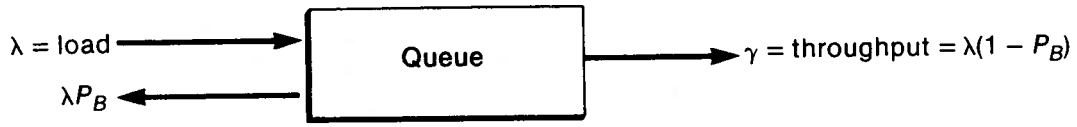

Figure 2-13 Relation between throughput and load

defined as the average number of arrivals/sec, is shown applied to a queue. With the probability of blocking given by P_B, the net arrival rate is then $\lambda(1 - P_B)$. But this must be the same as the throughput γ, or the number of customers served/sec for a conserving system. We thus have

$$\gamma = \lambda(1 - P_B) \tag{2-22}$$

as shown in Fig. 2-13. (A more detailed discussion, in the context of blocking in circuit-switched systems, appears in Chapter 10.)

One can calculate the throughput another way, however, by focusing on the output of the system. In particular, for a single-server queue, the type of queue under discussion, the average rate of service would be μ, in customers/sec served on the average, *if* the queue were always nonempty. Since the queue is sometimes empty, with probability p_0, the actual rate of service, or throughput γ, is less than μ. More precisely, $\gamma = \mu(1 - p_0)$, since $(1 - p_0)$ is the probability that the queue is nonempty. (So long as there is at least one customer in a queue, the average rate of service will be μ.) $(1 - p_0)$ is thus the (single-server) utilization. As a check, consider the infinite M/M/1 queue. There is no blocking in that case, and the throughput $\gamma = \lambda$, the average arrival rate. We must thus have $\lambda = \mu(1 - p_0)$, or $p_0 = 1 - \rho$, $\rho = \lambda/\mu$, just as found earlier! In the case of the finite M/M/1 queue, we equate the net arrival rate $\lambda(1 - P_B)$ to the average departure rate $\mu(1 - p_0)$ (see Fig. 2-14), to obtain

$$\gamma = \lambda(1 - P_B) = \mu(1 - p_0) \tag{2-23}$$

It is left to the reader to show, using Eq. (2-19) in Eq. (2-23), that the blocking

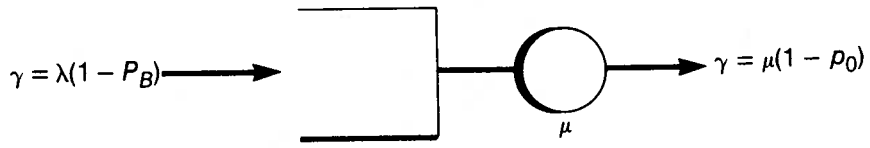

Figure 2-14 Throughput calculation, single-server queue

probability is in fact given by

$$P_B = p_N = (1 - \rho)\rho^N/(1 - \rho^{N+1}) \tag{2-24}$$

for the finite M/M/1 queue.

This equation for the blocking probability can be used for a simple design calculation. Specifically, what should the queue size N be to provide a prescribed blocking performance P_B? From Eq. (2-24) this also depends on ρ. For a small blocking probability, Eq. (2-24) may be simplified. With $\rho < 1$ and $N \gg 1$ we have

$$P_B \doteq (1 - \rho)\rho^N \qquad \rho^{N+1} \ll 1 \tag{2-25}$$

This is just the probability that the *infinite* queue is at state $n = N$ and indicates that for small P_B the probability that the finite queue is at state $n = N$ could just as well be calculated by assuming an infinite queue. Truncating the infinite queue at $n = N$ does not affect the queue statistics measurably *if* $\rho^N \ll 1$.

As examples of the use of Eq. (2-25), let $\rho = 0.5$. Then for $P_B = 10^{-3}$, $N \doteq 9$ customers that have to be accommodated. For a packet-switched network, the concentrator need only be capable of handling 9 packets. If $P_B = 10^{-6}$ is desired (one customer in 10^6 is rejected on the average), and $\rho = 0.5$, the number rises to $N \doteq 19$. Larger values of ρ (increased traffic) give rise to correspondingly larger values of N. As a simple example, say a concentrator in a packet-switched network handles packets that are 1200-bits long on the average. For a 2400-bps capacity line, the average transmission capacity is $\mu = 2$ packets/sec that can be delivered to the line. For $\rho = 0.5$, $\lambda = 1$ packet/sec is the allowable load on the concentrator. For a blocking probability of 10^{-3}, then, the line queue should be capable of accommodating 9 packets of average length 1200 bits. For $P_B = 10^{-6}$, this rises to 19 such packets.

Returning now to Eq. (2-20), the expression for the equilibrium state probability of the finite M/M/1 queue, we note that the condition $\rho < 1$ is no longer required for the equilibrium probabilities to exist. Since the queue is finite, the probabilities exist and are given by Eq. (2-20) for *all* values of ρ. Note in particular that as the load λ increases with respect to the capacity μ ($\rho = \lambda/\mu \to \infty$), the queue fills up more often and, in the limit of $\lambda \to \infty$, stays at state $n = N$ with a probability of 1. From Eq. (2-20), $p_n \to 0$, $n \neq N$, as $\rho \to \infty$, and $p_N \to 1$, $\rho \to \infty$. The region $\rho > 1$ is said to be the *congested* region; the higher queue states are more probable. The blocking probability $P_B = p_N$ approaches 1 as $\rho \to \infty$. This is shown in Fig. 2-15, which plots P_B as a function of the normalized load $\rho = \lambda/\mu$. It is readily shown, using L'Hôpital's rule, that $P_B = 1/(N + 1)$ at $\rho = 1$. For $N = 9$, as an example, $P_B \doteq 10^{-3}$ at $\rho = 0.5$, rises to $P_B \doteq 0.1$ at $\rho = 1$, is approximately 0.5 at $\rho = 2$, and continues to increase to 1 as $\rho \to \infty$. This indicates the queue is often blocked for ρ in the congested region.

The throughput of the queue, closely equal to the load λ for small ρ, eventu-

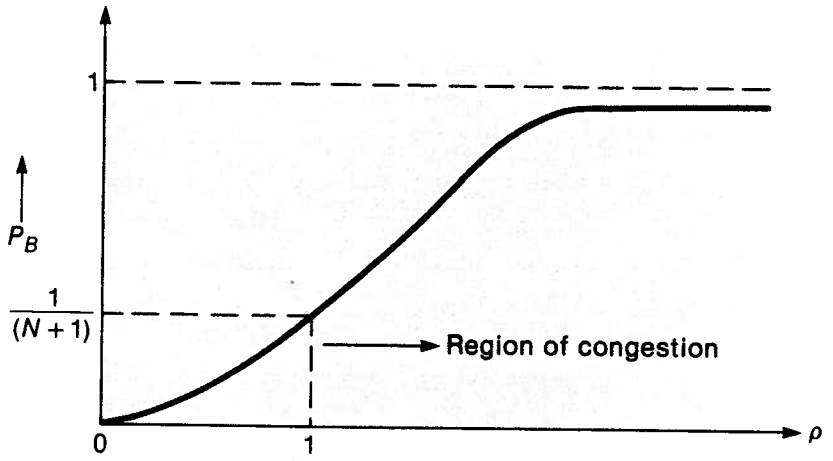

Figure 2-15 Blocking probability, finite M/M/1 queue, congested region

ally levels off and approaches the throughput capacity μ as ρ increases. The specific expression for the normalized throughput γ/μ as a function of the normalized load $\rho = \lambda/\mu$ is obtained by using Eq. (2-19) in Eq. (2-23). This gives

$$\gamma/\mu = (1 - p_0) = \rho(1 - \rho^N)/(1 - \rho^{N+1}) \qquad (2\text{-}26)$$

Equation (2-26) is sketched in Fig. 2-16. At $\rho = 1$, it is readily shown that $\gamma/\mu = N/(N+1)$ as indicated in Fig. 2-16. For $N = 9$, then, $\gamma/\mu = 0.9$, at

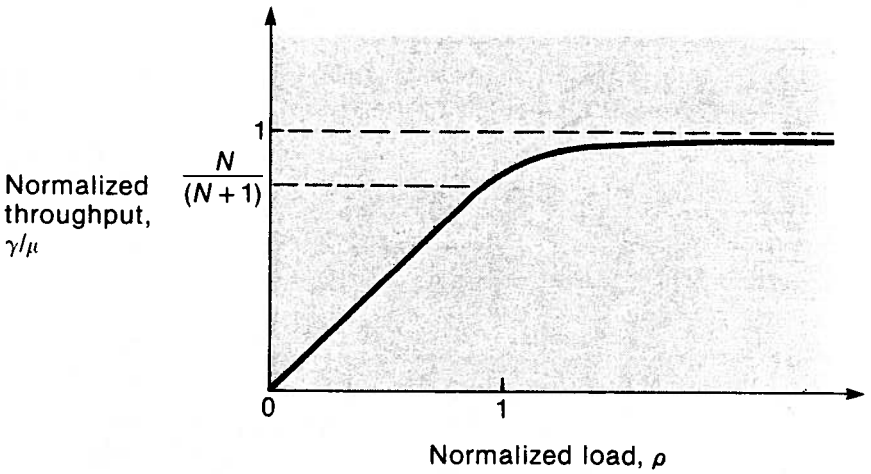

Figure 2-16 Throughput-load characteristic, finite M/M/1 queue

$\rho = 1$. Above this value of load, the normalized throughput levels off to values approaching 1 more and more closely.

We focus now on the uncongested region, $\rho < 1$, for which it is sufficient to use the infinite buffer analysis. From the state probabilities p_n, given by Eq. (2–18), one can calculate various statistics of interest. In particular, consider the average number of customers $E(n)$ (packets or calls queued) appearing in the queue, including the one in service. From the definition of the mean value of random variables, we have immediately

$$E(n) = \sum_{n=0}^{\infty} np_n = \rho/(1-\rho) \tag{2-27}$$

using Eq. (2–18) and carrying out the indicated summation. Equation (2–27) demonstrates the well-known queueing phenomenon that all of us have experienced in everyday life. When there is a relatively low load on the system ($\rho = \lambda/\mu \leq 0.5$, say), the average number of customers in the system is relatively small (<1 for $\rho < 0.5$). As ρ increases, approaching 1, the average number increases dramatically, rising because of the $(1-\rho)$ term in the denominator. In a real, finite queue system, the number would, of course, not shoot up as fast in the vicinity of $\rho < 1$, but Eq. (2–27) for the infinite queue does provide a good model for the finite queue case. Equation (2–27) is plotted in Fig. 2–17.

From a comparison of Figs. 2–15 to 2–17, one can make some statements about queueing performance that will be reiterated elsewhere in the book in discussing the performance of networks. As the load on the system increases,

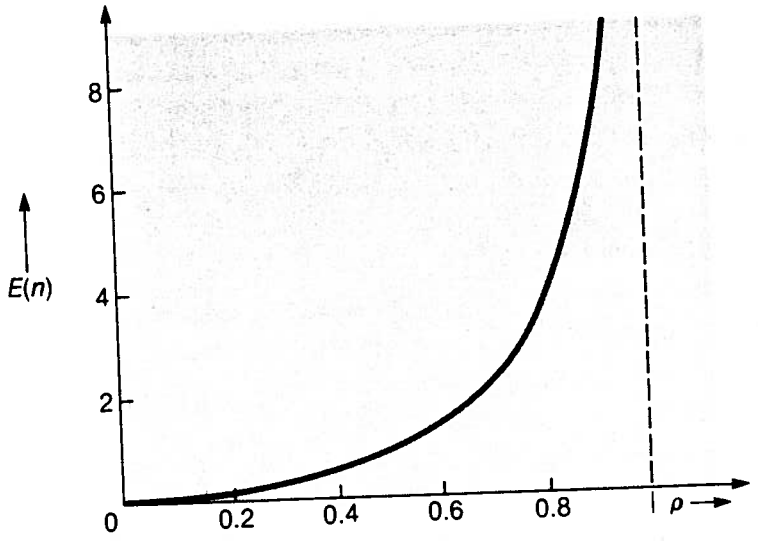

Figure 2–17 Average queue size, M/M/1 queue

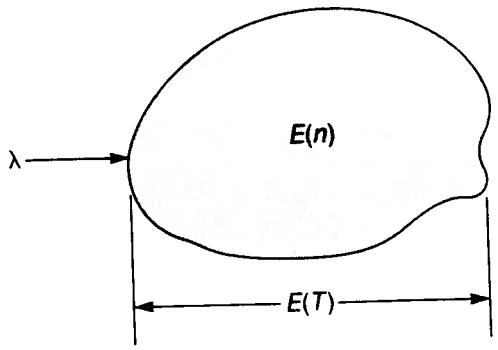

Figure 2-18 The environment of Little's formula

the throughput increases as well. More and more customers are blocked, however, and the average number of customers in the queue $E(n)$ also increases rapidly. This increase in $E(n)$ translates itself into increased time delay in the queue. There is thus a typical trade-off in performance: As the load increases the throughput goes up (a desirable characteristic) but blocking and time delay also increase (undesirable characteristics). In Chapter 5, in discussing congestion in more detail, we shall in fact see that queueing deadlocks can occur at high load. In this case, only apparent when two or more queues forming a network attempt to pass customers (packets) to one another, *nothing* will move, and the throughput drops to zero!

To find the time delay through the queue (this includes time spent waiting on the queue in addition to the service or transmission time), one invokes a simple formula that we shall be using throughout this book. The formula, called appropriately Little's formula after the individual who first proved it [LITT], says simply that a queueing system, with average arrival rate λ and mean time delay $E(T)$ through the system, has an average queue length $E(n)$ given by the expression

$$\lambda E(T) = E(n) \tag{2-28}$$

The relations among these three quantities are diagrammed in Fig. 2-18. We shall provide a proof of this expression in the next section. Suffice it to say here that the relation is very general and is valid for all types of queueing systems, including priority disciplines. The parameter λ is interpreted to be the arrival

[LITT] J. D. C. Little, "A Proof of the Queueing Formula $L = \lambda W$," *Operations Res.*, vol. 9, no. 3, 1961, 383-387.

rate *into* the system. It thus corresponds to our throughput γ. This should be apparent since customers that are turned away cannot contribute to delays in the system.

Applying Eq. (2-28) to the M/M/1 queue under discussion as shown in Fig. 2-19, one has immediately, using Eq. (2-27) for $E(n)$,

$$E(T) = E(n)/\lambda = 1/\mu/(1 - \rho) \tag{2-29}$$

This expression for the average time delay $E(T)$ through the M/M/1 queue has an interesting interpretation. For $\rho \ll 1$, $E(T) = 1/\mu$, exactly the average service time. This is the case, from Eq. (2-27), when there are few customers on the average in the queue. Hence very little time is spent waiting in the queueing system on the average, and the time delay is almost always due to service or transmission time. As the normalized load or traffic intensity increases, however, typical queueing behavior is experienced, with $E(T)$ beginning to rapidly increase. This is shown in Fig. 2-20 for $E(T)$ as a function of ρ. The normalized delay, $E(T)/1/\mu$—or $\mu E(T)$, the time delay relative to transmission time—is plotted in the figure. For $\rho = 0.5$, for example, the average delay doubles, to $2/\mu$. The average wait time in the queue at this point equals the average service time. For $\rho = 0.8$, the average delay is $5/\mu$, so that, on the average, there are $4/\mu$ units of wait time.

It is apparent that for the single-server queue the following simple relation between the average wait time $E(W)$ and the average delay $E(T)$ through the queue must hold:

$$E(T) = E(W) + 1/\mu \tag{2-30}$$

The connection between $E(T)$ and $E(W)$ is diagrammed in Fig. 2-19. Little's theorem enables us to find an explicit relation for the average number of

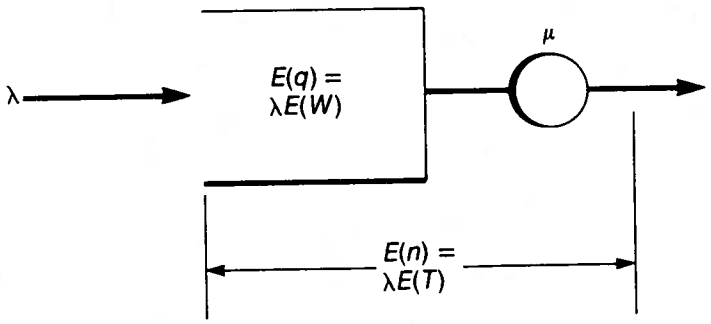

Figure 2-19 Little's formula applied to M/M/1 queue

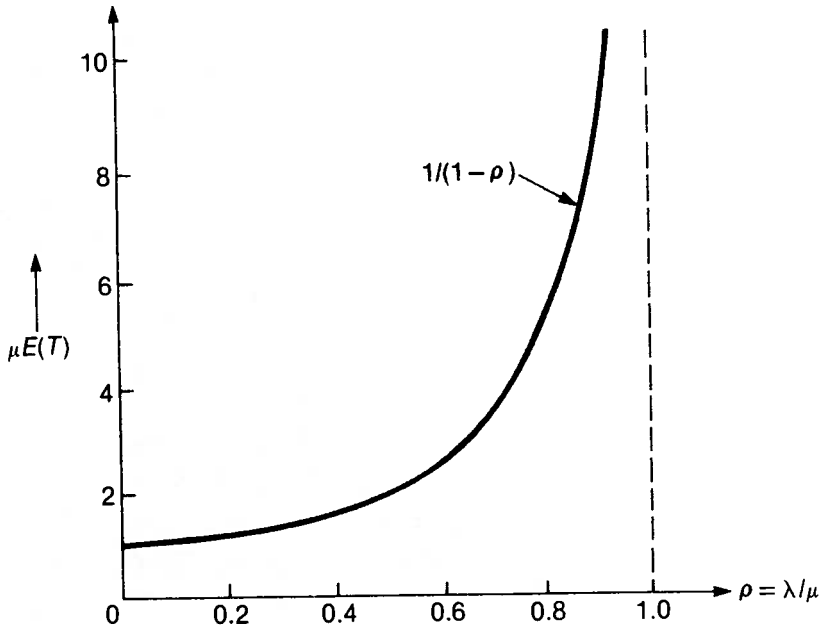

Figure 2-20 Normalized time delay, M/M/1 queue

customers $E(q)$ waiting in the queue (see Fig. 2-19). This must be given by

$$\begin{aligned} E(q) &= \lambda E(W) = \lambda E(T) - \lambda/\mu \\ &= E(n) - \rho \end{aligned} \tag{2-31}$$

(Recall that Little's formula is very general and that it can also be applied to a portion of a queueing system, as shown in Fig. 2-19.) As a check, ρ in Eq. (2-31) must represent the average number of customers in service. This is obviously less than one, since a customer is either in service or not. Focusing on the service station itself, there is a probability $p_0 = 1 - \rho$ that no one is in service (the queue is empty), hence a probability $(1 - p_0) = \rho$ that *one* customer is in service. The average is thus just ρ. The result for the M/M/1 queue, $p_0 = 1 - \rho$, may be generalized to *any* single-server queue.* From Eq. (2-31), ρ is always the average number in service. Hence $p_0 = (1 - \rho)$ must always be true.

Because of the great utility and generality of Little's formula, we devote the next section to a simple derivation.

* The queue must be *work conserving*, in the sense that if there is work to be done—i.e., a customer to be served—the customer will be served, and that a customer is not rejected once admitted into the system.

2–3 Little's Formula, $L = \lambda W$†

Consider a queueing system as shown in Fig. 2–21. For simplicity we consider the waiting section only. Let $A(t)$ represent the cumulative arrivals to the queue at time t, and let $D(t)$ represent the cumulative departures that go into service after waiting. Then $L(t) = A(t) - D(t)$ is the number waiting in the system at time t. No assumptions need be made about the arrival process or the departure process. We only stipulate that all customers entering the system are ultimately served (this is then a work-conserving system).

In particular, let customers arrive at time $t_j, j = 1, 2, \ldots$. $A(t)$ then represents the number of such arrival times, up to the time t. A typical state of the function $A(t)$ appears in Fig. 2–22. At each arrival, $A(t)$ increases by 1. Let the customers depart, moving to the service station, at times $t'_j \geq t_j$. To simplify the discussion initially, say that the service discipline is FIFO (the result obtained will be shown to be general, independent of this initial assumption). For this case, the departure times must increase monotonically, $t'_1 < t'_2 < t'_3 < \ldots$. A typical set of departure times appears in Fig. 2–22 with the corresponding curve for the cumulative departures $D(t)$ indicated as well. Note that departures coincide with arrivals when the system is empty. The wait times, W_j, for customers are also indicated. These obviously represent the time each customer spends in the system between arrival and departure. From Fig. 2–22 one can write, by inspection, simple relations between the different quantities shown that lead directly to Little's formula. Consider a starting time 0 and a later time τ, at both of which $A(t) = D(t)$. Examples in Fig. 2–22 include t'_2, t'_5, and t'_6. Let $n(\tau) = A(\tau) - A(0)$ be the number of arrivals in the interval τ. Then the mean arrival rate in the interval $(0,\tau)$ is just

$$\lambda(\tau) = n(\tau)/\tau \tag{2-32}$$

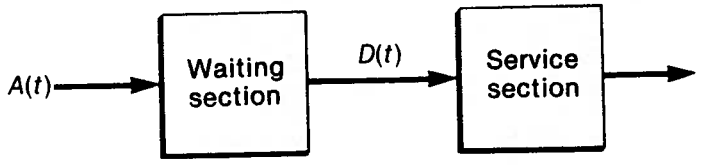

Figure 2–21 Derivation of Little's formula for a queueing system

† This section follows the approach of Kobayashi in [KOBA].

[KOBA] H. Kobayashi, *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology*, Addison-Wesley, Reading, Mass., 1978.

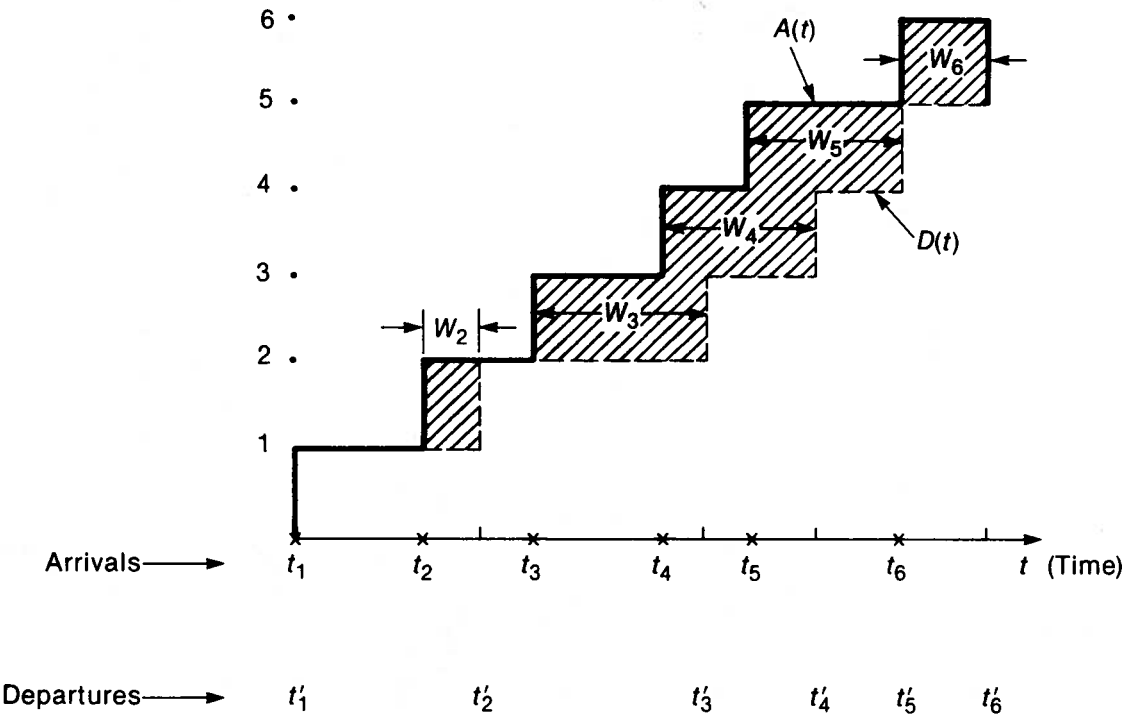

Figure 2-22 Arrivals and departures, FIFO (FCFS) queueing system

(An example in Fig. 2-22 would be six arrivals in the interval between t_1 and t_6.)

Focus now on the cross-hatched areas in Fig. 2-22. The cumulative area in the interval $(0,\tau)$ is just the function

$$\int_0^\tau L(t)dt$$

since $L(t) = A(t) - D(t)$, as noted earlier and as indicated in Fig. 2-22. Since this area is made up of a series of rectangles of unity height and width W_j as shown we obviously must have the equality

$$\sum_{j=1}^{n(\tau)} W_j = \int_0^\tau L(t)dt \tag{2-33}$$

But consider the quantity

$$\overline{W}(\tau) \equiv \sum_{j=1}^{n(\tau)} W_j / n(\tau) \tag{2-34}$$

This is the *average waiting time* in the interval $(0, \tau)$.

Consider also the expression

$$\bar{L}(\tau) \equiv \int_0^\tau L(t)dt/\tau \tag{2-35}$$

This must represent the *average number of customers* in the system in the interval $(0,\tau)$, since $L(t)$ is the number at time t. From Eq. (2-33), there is a close connection between $\bar{W}(\tau)$ and $\bar{L}(\tau)$. In particular, using Eqs. (2-34) and (2-35) in Eq. (2-33) and recalling the defining relation Eq. (2-32) for the arrival rate $\lambda(\tau)$, we have simply

$$n(\tau)\bar{W}(\tau)/\tau = \bar{L}(\tau) = \lambda(\tau)\bar{W}(\tau) \tag{2-36}$$

This is just Little's formula derived for the special case of FIFO service and over a finite interval $(0,\tau)$. Now let $\tau \to \infty$ and assume that the quantities of interest all approach definite limits:

$$\bar{W}(\tau) \to \bar{W},\ \lambda(\tau) \to \lambda,\ \bar{L}(\tau) \to \bar{L}$$

We then get Little's formula

$$\bar{L} = \lambda\bar{W} \tag{2-37}$$

with $\bar{L}$ the average number of customers in the queueing system, $\bar{W}$ the average waiting time, and λ the arrival rate. This result is easily extended to include the service station as well.

That Little's formula Eq. (2-37) holds generally for *any* service discipline is shown as follows. Consider

$$\sum_{j=1}^{n(\tau)} W_j = \sum_{j=1}^{n(\tau)} (t_j' - t_j)$$

from Fig. 2-22. This may be rewritten as

$$\sum_{j=1}^{n(\tau)} W_j = \sum_{j=1}^{n(\tau)} t_j' - \sum_{j=1}^{n(\tau)} t_j$$

Written in this form, it is apparent that $\sum_j W_j$ depends only on the sum of the departure times and not on the difference $t_j' - t_j$ used in the derivation assuming FIFO service. The individual departure times may depend on the service discipline, but Little's formula applied to any discipline holds nonetheless.

As an example, consider the last come-first served (LCFS) discipline. Typical plots for this case appear in Fig. 2-23. The reader is asked to demonstrate, using this figure, that the equality Eq. (2-33) holds here as well.

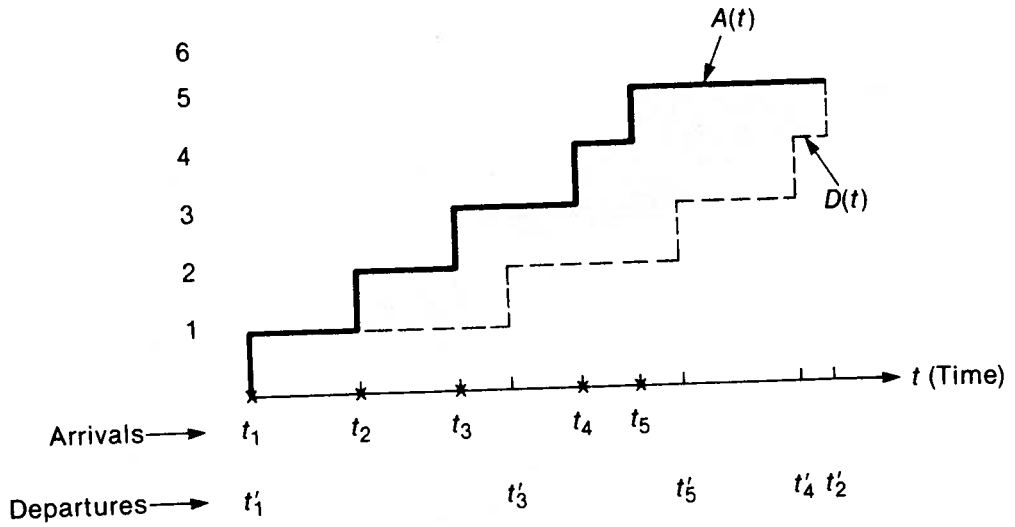

Figure 2-23 Arrivals and departures, LCFS discipline

2-4 State-dependent Queues: Birth-death Processes

The M/M/1 queue analysis carried out in detail earlier is readily extended to single-queue systems in which arrival and departure (service) rates are dependent on the state of the system. The processes are often called *birth-death processes* [COX]. We carry out a simple analysis of such a system in this section and provide some typical examples. A multiserver exponential queueing system, of the type M/M/m, will be found to fall in this class. A blocking system with no waiting room for customers, which often arises in the modeling of circuit-switched systems, will provide another example. Other examples appear in chapters that follow, particularly Chapter 5, on congestion control in packet-switched networks, and Chapter 10, on the traffic analysis of circuit-switched networks.

The state-dependent generalization of our previous M/M/1 model, with Poisson arrivals and exponential service distribution, is made for a system in state n simply by defining the probability of one arrival in the infinitesimal interval $(t, t + \Delta t)$ to be $\lambda_n \Delta t + 0(\Delta t)$, with the probability of no arrivals defined to be $(1 - \lambda_n \Delta t) + 0(\Delta t)$. The memoryless assumption is again invoked so that arrivals in the interval $(t, t + \Delta t)$ are independent of arrivals in other intervals. The arrival process is thus another example of a Markov process [COX], [PAPO]. A special case is our earlier Poisson process, with $\lambda_n = \lambda$, a constant,

independent of state. This arrival process is often called a birth process, since $\lambda_n \Delta t$ can be visualized as representing the probability of "birth" of a customer, given n customers already in the system. Note, as previously, that with this model at most *one* customer at a time can arrive in the system in the interval $(t, t + \Delta t)$.

In a similar manner, the state-dependent departure or death process is generalized from our previous Poisson departure process to be one for which the probability of departure of *one* customer in the infinitesimal interval $(t, t + \Delta t)$ is defined to be $\mu_n \Delta t + 0(\Delta t)$, given the number of customers in the system is n. The probability of departure of no customers is $(1 - \mu_n \Delta t) + 0(\Delta t)$, and the memoryless assumption is again invoked. The state-dependent departure or death process is similarly an example of a Markov process.

Combining these two processes, as was done in the case of the M/M/1 queue, again letting $\Delta t \to 0$ and taking the system to be in statistical equilibrium, the balance equation governing the operation of the combined birth-death process or state-dependent queueing system, at equilibrium (see Fig. 2-24), may be written by inspection:

$$(\lambda_n + \mu_n)p_n = \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} \tag{2-38}$$

Equation (2-15) for the M/M/1 queue is a special case. The corresponding state diagram appears in Fig. 2-25. The parameter λ_n represents the rate of arrival of a customer, given the system is in state n; p_n is the equilibrium probability that the system is in state n; μ_n is the rate of departure of a customer, given the system is in state n. The balance equation may again be obtained by equating the rate of departure *from* state n (the left-hand side of Eq. (2-38)) to the rate of arrival at state n (the right-hand side of Eq. (2-38)). Alternatively, from Fig. 2-25, one can obtain Eq. (2-38) by enclosing the state n with a closed surface, as was done in Fig. 2-11, and then equating the "flux" leaving the state to the "flux" entering it.

The same argument indicates that the solution to Eq. (2-38), extending the earlier M/M/1 analysis, is given by

$$\lambda_n p_n = \mu_{n+1}p_{n+1} \tag{2-39}$$

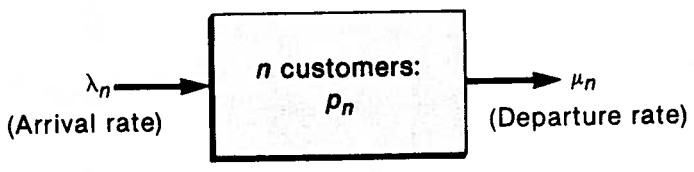

Figure 2-24 State-dependent queueing system

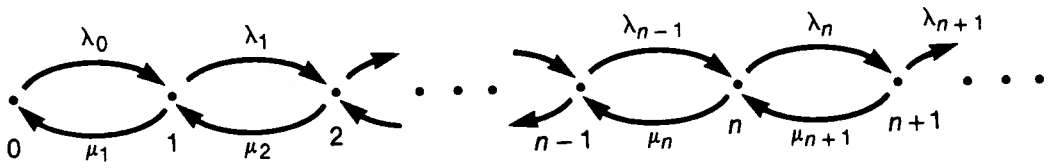

Figure 2-25 State diagram, state-dependent queue (birth-death process)

This is to be compared with the earlier equation, Eq. (2-16), representing the solution for the M/M/1 queue with state-independent Poisson arrival and departure processes. Using Eq. (2-39), it is left to the reader to show that the equilibrium probability p_n for the birth-death process, or state-dependent queueing system, is given by

$$p_n/p_0 = \prod_{i=0}^{n-1} \lambda_i / \prod_{i=1}^{n} \mu_i \qquad (2-40)$$

The unknown state probability p_0 for a finite queue holding at most N customers is again found by invoking the normalization condition $\sum_{n=0}^{N} p_n = 1$. Such a queueing system will *always* be stable. For the infinite queue ($N \to \infty$), stability is again ensured by having $p_0 > 0$.

Some examples of the application of these results are of interest. Consider first the case of two outgoing trunks (transmission links) connecting a statistical concentrator or packet switch to a neighboring packet switch, or node, in a packet-switched network. Data packets use either one of these two trunks randomly. What effect does adding a second trunk have on the operation of the system? Assume that packets arriving at the output queue driving the double transmission-link facility obey a Poisson process with average rate λ. Packets are again assumed to be exponentially distributed in length, with an average length $1/\mu$ in sec. The model for the resultant queueing system is then the one in Fig. 2-26. It is precisely that of an M/M/2 queue.

Consider the operation of this system now. If only one packet is available for transmission, it is immediately serviced by either trunk, at the service rate μ. If two or more packets are available, both trunks are occupied. Because of the exponential service length assumption made, the probability of *either* trunk completing service in the interval $(t, t + \Delta t)$ is $\mu\Delta t$, and the probability of *one* completion in the same interval is $2\mu\Delta t$. (Under the exponential assumption the probability of *both* trunks completing a transmission in the same infinitesimal interval is $0(\Delta t)$, and hence goes to zero as $\Delta t \to 0$). But this system is precisely that of a birth-death system, with $\lambda_n = \lambda$, independent of n, and $\mu_n = \mu$, $n = 1$,

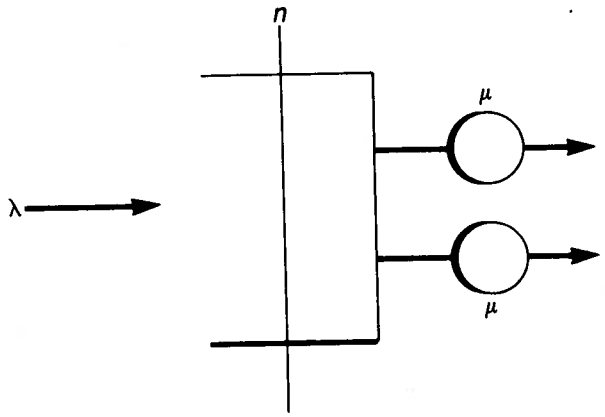

Figure 2-26 M/M/2 queue

$\mu_n = 2\mu$, $n \geq 2$. For the M/M/2 queue, then, one has, from Eq. (2-40),

$$p_n/p_0 = (\lambda/2\mu)^{n-1}\left(\frac{\lambda}{\mu}\right) = 2\rho^n, \qquad n \geq 1 \qquad \rho \equiv \lambda/2\mu \tag{2-41}$$

The parameter ρ has been defined in terms of 2μ here, since the effective service rate is 2μ for the state $n \geq 2$. With at least two packets in the queue, the system serves at twice the rate of a single-trunk system, reducing the effective traffic intensity correspondingly, and hence acting to reduce queue congestion (as measured by the time delay) as well. These qualitative considerations will be borne out quantitatively. Allowing the queue to be an infinite one for simplicity, one finds, from the normalization condition, that

$$p_0 = (1-\rho)/(1+\rho) \qquad \rho \equiv \lambda/2\mu \tag{2-42}$$

and

$$p_n = \frac{2(1-\rho)}{(1+\rho)}\rho^n \qquad n \geq 1 \tag{2-43}$$

The average queue occupancy is readily shown to be given by

$$E(n) = \sum_{n=0}^{\infty} np_n = \frac{2\rho}{(1-\rho^2)} \qquad \rho = \lambda/2\mu \tag{2-44}$$

This is always less than the average queue occupancy for the M/M/1 case, $E(n)|_{M/M/1} = \rho/(1-\rho)$, $\rho \equiv \lambda/\mu$, as expected. The average time delay in the queue, wait time plus service time, is readily obtained using Little's formula:

$$E(T) = E(n)/\lambda = 1/\mu/(1-\rho^2) \qquad \rho \equiv \lambda/2\mu \tag{2-45}$$

This is, of course, always less than the M/M/1 result. Note in addition that because of the 2μ service rate for the state $n > 1$, the M/M/2 queue can operate out to twice the arrival rate of the M/M/1 queue: $\lambda < 2\mu$. Adding the additional server thus improves both the time delay and the throughput performance. This is diagrammed in Fig. 2-27, which shows the normalized time delay $\mu E(T)$ plotted versus the normalized arrival rate $\lambda/2\mu$. Also shown is a time-delay load curve for an M/M/1 queue with twice the service capacity, 2μ. The performance of this system is always better than that of either of the other two systems. In terms of performance, it is always more effective to double the transmission capacity than to add a second trunk at the original capacity, *if justified by cost considerations* or required by reliability considerations. The reason is obvious: The packet transmission time is halved, so that at low utilization ($\rho \ll 1$) more packets are being served per unit time. For at least two packets in the system, the probability of completion is the same in the two systems (the M/M/2 queue with service rate μ and the M/M/1 queue with service rate 2μ). As the traffic utilization increases, the average time delays of the two strategies (the one adds a second server, the other doubles the service rate) approach one another.

Consider now, in the M/M/2 case, the significance of the traffic intensity parameter ρ, defined here as $\lambda/2\mu$. As in the case of the M/M/1 queue, one argues that the maximum possible throughput is 2μ. This is not attained here

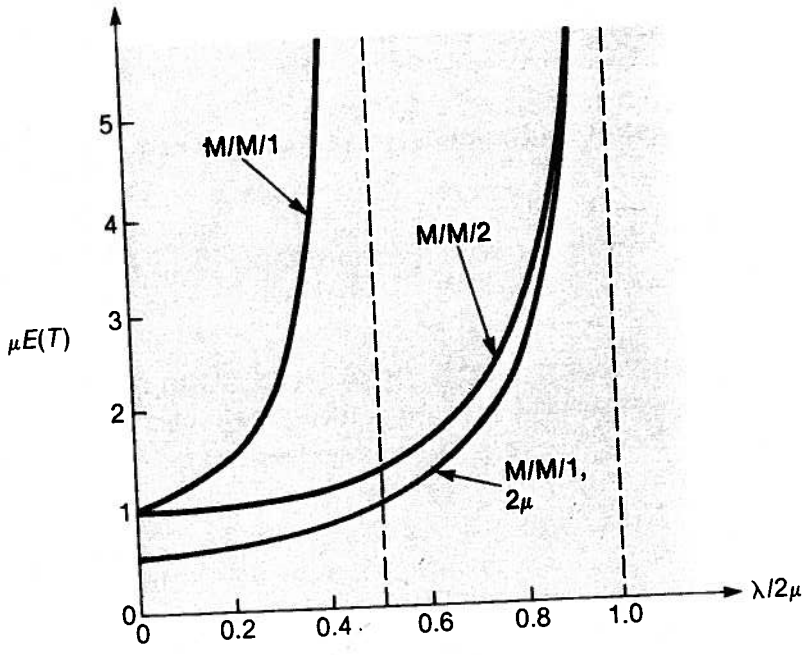

Figure 2-27 Performance characteristic, M/M/2 queue

because, with probability p_0, the queue may be empty; with probability p_1, only one server is utilized. The average throughput for the two-server case is therefore just

$$\gamma = \mu p_1 + 2\mu(1 - p_0 - p_1) \tag{2-46}$$

For the infinite M/M/2 queue, with no blocking, this should be just the average arrival rate or load on the system λ. Introducing the values for p_0 and p_1 for the infinite queue case from Eq. (2-42) and Eq. (2-43), respectively, one finds this to be precisely the case. This is the reason for introducing the parameter $\rho = \lambda/2\mu$ to represent the ratio of load to maximum transmission capacity of the system.

The second and third examples we discuss can be considered together. The second example extends the M/M/2 case just treated to the case in which a server is made available to any customer entering the system. In both the packet-switched and the circuit-switched cases this implies that the number of transmission links or trunks is always equal to the number of packets or calls desiring transmission. Thus there is never any queueing up for service or any possibility of blocking. The model for this is simply $\mu_n = n\mu$, for all n, if exponential service is again assumed. We again take the arrival rate to be Poisson, with average rate λ. For the infinite queue case, the queue structure one gets is then called the M/M/∞ queue. For this case, from Eq. (2-40), one finds quite readily that

$$p_n/p_0 = (\lambda/\mu)^n/n! \tag{2-47}$$

and that

$$p_0 = e^{-\rho} \qquad \rho \equiv \lambda/\mu \tag{2-48}$$

The probability of state occupancy is, in this case, given by the Poisson distribution.

The third example is called a "queue with discouragement." It models a system with customer flow control at the input. Specifically, for this model we let the state-dependent arrival rate λ_n be given by $\lambda_n = \lambda/(n+1)$, with λ a known constant. Only a single server is available in this case, so that $\mu_n = \mu$, for all n. This example could be used to model moviegoers or shoppers who arrive at the movie theatre or supermarket and find only a single line to serve them. When the line becomes too long (n is large), they are discouraged and turn away. In the context of packet transmission the arrival model represents one in which the maximum possible arrival rate is λ. As the queue length n increases, a system controller discourages packet arrivals (either by blocking or by shunting some arrivals elsewhere), so that the actual arrival rate decreases as n increases. (Flow control in packet networks will be described in Chapter 5.) For this example, a little thought will indicate that the probability p_n of queue occupancy is precisely

that given by Eq. (2-47). For an infinite queue, p_0 is given by Eq. (2-48) as well. The two examples, the M/M/∞ queue and the queue with discouragement, thus have the same solution for the probability of state. The average state occupancy is given in both cases by

$$E(n) = \sum_{n=0}^{\infty} np_n = \rho = \lambda/\mu \tag{2-49}$$

either by using Eqs. (2-47) and (2-48) and carrying out the summation indicated or by invoking the property of the Poisson distribution noted earlier (Eq. (2-2)). This average number of customers is always less than $\rho/(1-\rho)$, the average number in the M/M/1 model, showing the benefit to be derived by either increasing the number of servers or controlling the input arrival rate.

The two examples do differ, however, in their time delay-throughput characteristics. This should be the case since, despite the identical solution for the probability of state in the two cases, they do represent different physical situations. Take the M/M/∞ queue first. Its throughput is just λ, the applied load, since $\lambda_n = \lambda$ for all values of n. From Little's formula, then, the average time delay is

$$E(T) = E(n)/\lambda = 1/\mu \tag{2-50}$$

using Eq. (2-49). But this is precisely the result expected since, with the number of servers always equal to the number of customers (packets or calls) in the system, there is no queueing, and the time delay is just the average service or transmission time $1/\mu$. As a check, consider the throughput γ, as calculated at the system *output*. For a state-dependent departure (death) process this must be the average departure rate, found by averaging over all the departure rates, μ_n, for all n. Specifically, in this case, with $\mu_n = n\mu$, we have

$$\gamma = \sum_{n=0}^{\infty} \mu_n p_n = \mu \sum_{n=0}^{\infty} np_n = \mu E(n) = \lambda \tag{2-51}$$

invoking Eq. (2-49) to obtain the last result.

Consider the third example now, that of the queue with discouragement. Here, since the input arrival rate is state dependent, one must average over all the states to find the average arrival rate or throughput of the system. We thus have

$$\gamma = \sum_{n=0}^{\infty} \lambda_n p_n = \mu(1 - e^{-\rho}) \qquad \rho = \lambda/\mu \tag{2-52}$$

using the Poisson distribution of Eqs. (2-47) and (2-48) for p_n, the definition of λ_n in this case, and carrying out the indicated summation. In this case the throughput could have been obtained much more simply, as indicated by the

form of the result in Eq. (2-52), by recalling that for a single-server queue of capacity μ, the throughput is just $\mu(1 - p_0)$. From Eq. (2-48) one immediately obtains Eq. (2-52).

Using Eqs. (2-49) and (2-52), one now finds the normalized average time delay in the system to be given by

$$\mu E(T) = \mu E(n)/\gamma = \rho/(1 - e^{-\rho}) \qquad \rho = \lambda/\mu \tag{2-53}$$

For small ρ ($\rho \ll 1$), the time delay is just $1/\mu$, the average transmission or service time, as expected. The throughput is just $\gamma = \lambda$ in this case, from Eq. (2-52). As the parameter $\rho = \lambda/\mu$ increases, the throughput approaches the maximum value μ, while the normalized time delay continues to increase linearly with ρ, reflecting the fact that the average queue occupancy $E(n) = \rho$, a linear increase with ρ. No limit is required on ρ in this case to ensure stability, since the flow (discouragement) control serves to keep the system stable even though ρ increases indefinitely. Alternatively stated, the Poisson distribution form for p_n, given by Eqs. (2-47) and (2-48), ensures that $p_0 > 0$ for all ρ, and the system remains stable as ρ increases. Equation (2-47) implies that higher states become more probable as ρ increases, with a corresponding increase in the average state occupancy $E(n)$.

The fourth example is a special case of the M/M/∞ queue, with a finite number of servers and no waiting room. Specifically, let $\mu_n = n\mu$, $1 \le n \le N$, $\lambda_n = \lambda$, and block all arrivals if $n = N$. This is often written as an M/M/N/N system. It models a system in which customers arrive according to a Poisson arrival process with average rate λ and always find a trunk (transmission link) available until a maximum number of trunks is occupied. At this point customers arriving are blocked. This model, with the addition of the exponential service time assumption, has been used for many years as a basic design model for telephone exchanges; it will be discussed and used in detail in Chapter 10. A conceptual diagram appears in Fig. 2-28. We focus on the circuit-switched (telephone) terminology. N servers (trunks or transmission links) are available to handle calls. When all trunks are occupied, further calls arriving are blocked. Since there is no queueing (no waiting room) allowed in this system, it is referred to as a pure loss system. The performance parameter of interest in the circuit-switched (telephone) application is the probability of blocking P_B. The solution here is readily obtained. Since $\mu_n = n\mu$ and $\lambda_n = \lambda$, one finds, using Eq. (2-40) and invoking the normalization condition

$$\sum_{n=0}^{N} p_n = 1$$

that

$$p_n = \frac{\rho^n/n!}{\sum_{\ell=0}^{N} \rho^\ell/\ell!} \qquad \rho = \lambda/\mu \tag{2-54}$$

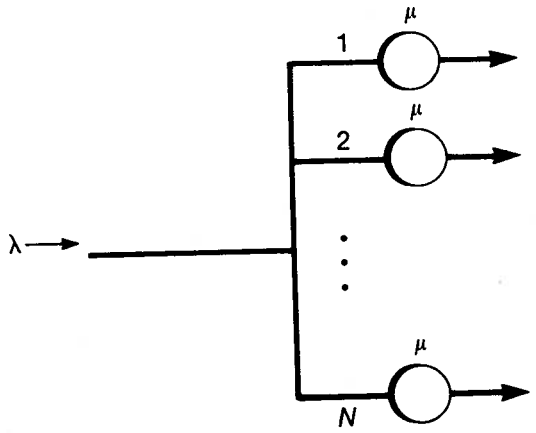

Figure 2-28 M/M/N/N system: no waiting room

In particular, blocking occurs with $n = N$, so that the blocking probability is given by

$$P_B = \frac{\rho^N/N!}{\sum_{\ell=0}^{N} \rho^\ell/\ell!} \qquad \rho = \lambda/\mu \tag{2-55}$$

This equation for the blocking probability is often called the Erlang-B distribution, Erlang distribution of the first kind, or the Erlang loss formula, after the great Swedish engineer A. K. Erlang, who first studied the traffic performance of telephone systems using a statistical approach in the early part of the twentieth century. We shall refer to this equation in detail in Chapter 10 in our discussion of circuit-switched traffic analysis. There we shall adopt the symbol A in place of ρ to represent the total normalized load or traffic intensity λ/μ on the system. The units of A are given in terms of Erlangs.

It is left for the reader to show that the average number of calls in the system is

$$E(n) = \rho(1 - P_B) \qquad \rho = \lambda/\mu \tag{2-56}$$

As the traffic intensity ρ increases, $P_B \to 1$ and $E(n) \to N$. The throughput γ, in calls per unit time accepted by the system and hence delivered at the output, is again

$$\gamma = \lambda(1 - P_B) \tag{2-57}$$

Averaging over the state-dependent service rate at the output of the system, one also finds

$$\gamma = \sum_{n=0}^{N} \mu_n p_n = \mu E(n) \tag{2-58}$$

invoking the definitions of μ_n and of $E(n)$, respectively. As the traffic builds up, the throughput approaches its maximum value of $N\mu$. This corresponds to the case $\rho = \lambda/\mu \gg N$, in which situation most of the calls arriving are being blocked as well; thus $P_B \rightarrow 1$. The average delay through the system, for those calls accepted, is just $1/\mu$, the service time (called the holding time in telephone practice). As a check, we have, invoking Little's formula,

$$E(T) = E(n)/\gamma = 1/\mu \tag{2-59}$$

from Eq. (2-58).

Other examples of state-dependent queue analysis will be encountered throughout the book, in analyzing quantitatively the performance of packet-switched, circuit-switched, and integrated networks.

2-5 M/G/1 Queue: Mean Value Analysis

In the previous section we extended the M/M/1 queue analysis to the case of state-dependent arrival and service times (the birth-death process). The state-dependent model is extremely useful and is frequently used, as indicated by some of the examples. However, it still relies on the Markov memoryless property for both the arrival process and the service-time distribution. In this section we extend the analysis to one other case, that of a *general* service-time distribution. Packets or calls may thus have an arbitrary (but known) length or service distribution. However, the arrival process will be taken to be Poisson, a single server is assumed, and the queue buffer size (waiting room) is taken to be infinite. Such a queue is called an M/G/1 queue, using the Kendall notation, with G obviously standing for general service distribution.

For simplicity, in this section we shall focus on *average* (mean) occupancy and *average* time delay only. More general discussions appear in books on queueing theory. The book by L. Kleinrock is a particularly good example [KLEI 1975a]. A brief discussion appears in [SCHW 1977], Chapter 6. We shall show that the average queue occupancy $E(n)$ and the average time delay through the queue $E(T)$ are given, respectively, by the following rather simple-looking expressions:

$$E(n) = \left(\frac{\rho}{1-\rho}\right)\left[1 - \frac{\rho}{2}(1 - \mu^2\sigma^2)\right] \tag{2-60}$$

[KLEI 1975a] L. Kleinrock, *Queueing Systems. Volume 1: Theory*, John Wiley & Sons, New York, 1975.

[SCHW 1977] M. Schwartz, *Computer-Communication Network Design and Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1977.

and

$$E(T) = \frac{E(n)}{\lambda} = \frac{1/\mu}{(1-\rho)}\left[1 - \frac{\rho}{2}(1 - \mu^2\sigma^2)\right] \qquad (2-61)$$

These are called the Pollaczek-Khinchine formulas after two Russian mathematicians. The parameter ρ is again given by $\lambda/\mu = \lambda E(\tau)$, with λ the average (Poisson) arrival rate and $E(\tau) = 1/\mu$ the average service time. The parameter σ^2 is the variance of the service-time distribution.

Note that both expressions appear closely related to the corresponding results in the M/M/1 case. (These are the leading terms, before the brackets, in both equations.) This is a remarkable result: The average queue occupancy (and corresponding time delay) for a queue with Poisson arrivals and *any* service-time distribution is given by the result obtained for an exponential service distribution with the same average service time, multiplied by a correction factor. This correction factor, the term in brackets in Eqs. (2-60) and (2-61), is seen to depend on the ratio of the variance σ^2 of the service distribution to the average value squared, $1/\mu^2$.

Recall that the variance of the exponential distribution is $\sigma^2 = 1/\mu^2$, i.e., the square of the average value. Setting $\sigma^2 = 1/\mu^2$ in Eqs. (2-60) and (2-61), then, one obtains the results derived earlier for the M/M/1 queue. As σ^2 increases, with $\sigma^2 > 1/\mu^2$, the corresponding average queue occupancy and time delay increase as well. For $\sigma^2 < 1/\mu^2$, on the other hand, the average queue occupancy and time delay decrease relative to the M/M/1 result. As a special case, let all customers (packets or calls) have the *same* service length $1/\mu$. Then for $\sigma^2 = 0$,

$$E(n) = \frac{\rho}{(1-\rho)}\left(1 - \frac{\rho}{2}\right) \qquad \sigma^2 = 0 \qquad (2-62)$$

and

$$E(T) = \frac{1/\mu}{(1-\rho)}\left(1 - \frac{\rho}{2}\right) \qquad \sigma^2 = 0 \qquad (2-63)$$

A queue of this type, with fixed customer service time, is called an M/D/1 queue. The letter D represents *deterministic* service time. This is then a special case of the M/G/1 queue, with the smallest possible queue occupancy and delay. Note that for ρ not too large, $E(n)$ and $E(T)$ may be obtained (conservatively) by using the M/M/1 results. For $\rho \to 1$, the M/D/1 results differ by 50 percent from the M/M/1 values.

The interesting thing is that for the general M/G/1 results of Eqs. (2-60) and (2-61) the dominant behavior of average queue occupancy and time delay is always the $1/(1-\rho)$ term of the denominators. *All* infinite buffer queues, no matter what the service distribution, thus tend to exhibit the same queue-congestion behavior as $\rho = \lambda/\mu \to 1$. Those with larger variance in their service distribution produce larger queue occupancy and time delay, on the average.

This is to be expected since the increased variance means a large spread in the service times, with a correspondingly higher probability that *longer* service times, leading to more congestion, will be encountered.

Equation (2–61) may be used to obtain a compact, general form for the average wait time $E(W)$ on the queue. Specifically, recall that $E(T)$ and $E(W)$ are related by the expression

$$E(W) = E(T) - 1/\mu \tag{2–64}$$

Substituting Eq. (2–61) for $E(T)$ into Eq. (2–64) and simplifying, one obtains the following simple result for the average wait time $E(W)$ in an M/G/1 queue:

$$E(W) = \frac{\lambda E(\tau^2)}{2(1-\rho)} \tag{2–65}$$

The term $E(\tau^2)$ is the second moment of the service-time distribution, given by

$$E(\tau^2) = \sigma^2 + 1/\mu^2$$

This is obviously a simpler expression to remember than the Pollaczek-Khinchine form of Eq. (2–61). It will be used in the next section in discussing the wait time in priority queues.

The derivation of Eq. (2–60) for the average queue occupancy $E(n)$ proceeds in a manner different from that used in the M/M/1 analysis. Because of the condition of general service-time statistics, with the corresponding lack of a memoryless property for service departures, one can no longer set up a simple balance equation for the states of the queue. Instead we use an approach that focuses on the change in buffer (queue) occupancy at the conclusion of service, i.e., the departure time. Specifically, label the time at which the jth customer departs the queue by the number j. The number of customers (packets or calls) remaining in the queue just after the departure is represented by the number n_j. A timing diagram indicating these quantities appears in Fig. 2–29.

A simple relation may be written connecting the number of customers in the queue at time j to the number at time $(j-1)$. Let v_j be the number of customers arriving during the service interval of the jth customer. Then n_j and n_{j-1} are obviously connected by the relation

$$\begin{aligned} n_j &= (n_{j-1} - 1) + v_j \qquad n_{j-1} \geq 1 \\ &= v_j \qquad\qquad\qquad n_{j-1} = 0 \end{aligned} \tag{2–66}*$$

Equation (2–66), representing the queue dynamics for a general service-time distribution, enables us to find the steady-state statistics of the queue *at service*

* If the queue empties at $(j-1)$, we wait until the next customer arrives and in turn completes service. During this service v_j customers may arrive.

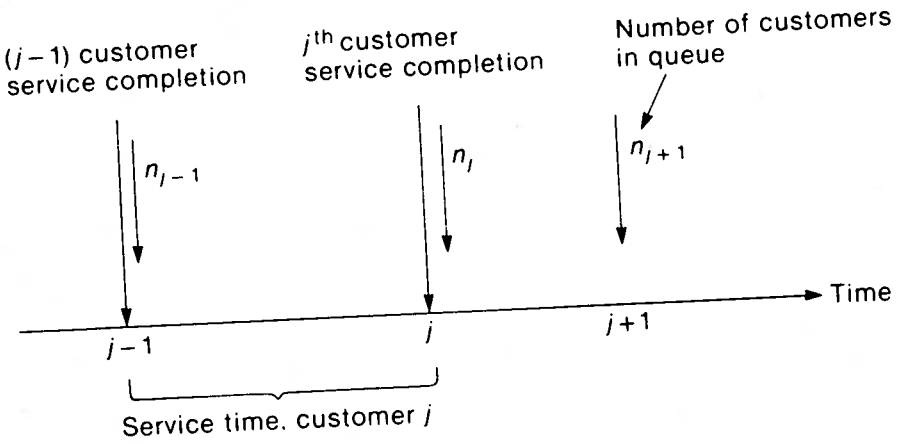

Figure 2-29 Timing diagram, general service-time distribution

completion times. For the special case of Poisson arrivals, on which we focus shortly, one can argue, because of the memoryless property of the arrivals, that the statistics found are the same at all points on the time axis.

Equation (2-66) may be rewritten in the following compact form:

$$n_j = n_{j-1} - u(n_{j-1}) + v_j \tag{2-66a}$$

where $u(x)$ is the unit step function defined as

$$\begin{aligned} u(x) &= 1 \qquad x \geq 1 \\ &= 0 \qquad x \leq 0 \end{aligned}$$

Now let the time $j \to \infty$ and assume that equilibrium has set in. Taking expectations on both sides of Eq. (2-66a), and noting that $E(n_j) = E(n_{j-1}), j \to \infty$, we get

$$E(v) = E[u(n)] \tag{2-67}$$

$E(v)$, the average number of arrivals in a service interval, will be expected to be less than one for an infinite queue in order to maintain a stable queue, as required for equilibrium. We shall show shortly that $E(v) < 1$, and in fact corresponds to the utilization ρ for a single-server queue, first encountered in the M/M/1 case. To demonstrate this, we note from the definition of the unit step function $u(n)$ that

$$E[u(n)] = \sum_{n=1}^{\infty} p_n = \text{prob. } (n > 0) \equiv \rho \tag{2-68}$$

Here p_n is, as throughout this chapter, the probability that the queue is in state n. From Eq. (2-68) we thus have $p_0 = (1 - \rho) > 0$, as found first in the M/M/1 case.

Equations (2-66) and (2-66a) may be used to obtain the state probabilities of the M/G/1 queue, using transforms or moment-generating functions [KLEI 1975a], [SCHW 1977]. For our purposes, as noted earlier, it suffices to find average queue occupancy only. To do this, we employ a trick. Square the left- and right-hand sides of Eq. (2-66a), take expectations, and again let $j \to \infty$. Noting that $E[u^2(n)] = E[u(n)] = E(v) = \rho$, that $E[nu(n)] = E(n)$, and assuming v_j and n_{j-1} to be *independent* (as is the case for Poisson arrivals), one obtains the following interesting result after some simplification:

$$E(n) = \frac{\rho}{2} + \frac{\sigma_v^2}{2(1-\rho)} \tag{2-69}$$

We have used $E(v) = \rho$ here; σ_v^2, yet to be found, is the variance of the number of customers arriving in a service interval.

To proceed, i.e., to find σ_v^2 specifically, we now invoke the Poisson arrival assumption. We also let the (general) service-time probability density function be given by $f_\tau(\tau)$, with the parameter τ representing the service time. To find σ_v^2 and $E(v)$ we must know the probability $P(v = k)$ that exactly k customers arrive in a service interval. For example, it is apparent from their definitions that

$$E(v) = \sum_{k=0}^{\infty} kP(v = k) \tag{2-70}$$

and that

$$\sigma_v^2 = \sum_{k=0}^{\infty} [k - E(v)]^2 P(v = k) \tag{2-71}$$

Defining the conditional probability $P(v = k|\tau)$ as the probability of k arrivals in τ sec, $P(v = k)$ is obviously given by

$$P(v = k) = \int_0^{\infty} P(v = k|\tau) f_\tau(\tau) d\tau \tag{2-72}$$

Focusing now on the M/G/1 queue case, for Poisson arrivals we have

$$P(v = k|\tau) = (\lambda\tau)^k e^{-\lambda\tau}/k! \tag{2-73}$$

Inserting Eq. (2-73) into Eq. (2-72), and then into Eq. (2-70), one finds after interchanging integration and summation that

$$E(v) = \lambda E(\tau) \equiv \rho \tag{2-74}$$

$$E(\tau) = \int_0^{\infty} \tau f_\tau(\tau) d\tau \tag{2-75}$$

The definition of $E(v)$, the average number of arrivals in a service interval, as the

utilization ρ is thus validated another way: Eq. (2–74) indicates that $E(v)$ is, for Poisson arrivals, given by the average customer arrival rate λ times the average service time $E(\tau)$, extending the definition used in the M/M/1 case.

The calculation of σ_v^2 for the case of Poisson arrivals proceeds in a manner similar to that of the calculation of $E(\tau)$. Again inserting Eq. (2–73) into Eq. (2–72) and then inserting the result into Eq. (2–71) (the defining equation for σ_v^2), interchanging summation and integration, and then simplifying the result, one finds that

$$\sigma_v^2 = \rho + \lambda^2\sigma^2 \tag{2–76}$$

with σ^2 the variance of the service-time distribution. In deriving this result, use is made of the fact that the variance of the Poisson distribution equals its mean value. Details are left to the reader.

Introducing Eq. (2–76) into Eq. (2–69), one obtains the desired Pollaczek-Khinchine form, Eq. (2–60). Equation (2–61) for the average time delay (wait time plus transmission time) through the queue follows directly from Little's theorem. Alternatively, a much simpler form of the Pollaczek-Khinchine formula results if one focuses on the average wait time $E(W)$ in the queue. This was noted earlier and the simple, compact form Eq. (2–65) for $E(W)$ was derived from Eq. (2–61) for the time delay $E(T)$ through the queue. Equation (2–65) will be found very useful in the next section in calculating the wait time in nonpreemptive priority queueing systems.

2–6 Nonpreemptive Priority Queueing Systems

The need to provide priority to certain classes of customers in a queueing system arises in many applications. Priority classes are used in many computer systems, in the computer control of digital switching exchanges, for deadlock prevention in packet-switching networks, and so forth. A simple example taken from packet switching serves to provide the necessary motivation at this point. Consider a packet-switching network that transmits, in addition to the normal data packets, control packets of much shorter length that carry out the vital operations of signaling, congestion notification, fault notification, routing change information, and so on. (Some of these control functions will be explored in the chapters following.) It is vital in many cases to ensure rapid distribution of these control messages. Yet without establishing a priority, they could easily queue up behind much longer data packets, delaying their arrival at the necessary destination points.

As a specific example, consider a network connected by 9600 bps transmis-

sion links. We designate data packets by the label 2, control packets by the label 1. Data packets are on the average 960 bits long, or $1/\mu_2 = 0.1$ sec, with a variance $\sigma_2^2 = 2(1/\mu_2)^2$, or $E(\tau^2) = 3(1/\mu_2)^2$. Control packets, on the other hand, are all 48 bits long, so that $1/\mu_1 = 5$ msec. In this case of fixed packet lengths $\sigma_1^2 = 0$. Focus on a single queue, served FIFO, that drives the outgoing transmission link. Let 20 percent of the total traffic be due to the short control packets, 80 percent to the much longer data packets. We thus have $\lambda_1 = 0.2\lambda$, $\lambda_2 = 0.8\lambda$, with λ the composite arrival rate at the queue in packets/sec. It is apparent that without priority the queueing of the combined traffic stream can be modeled as an M/G/1 queue. The combined traffic intensity is $\rho = \rho_1 + \rho_2$. Since packets of either kind arrive randomly, with rates λ_1 and λ_2, respectively, the second moment of the composite stream is given by the weighted sum of the second moments:

$$E(\tau^2) = \frac{\lambda_1}{\lambda} E(\tau_1^2) + \frac{\lambda_2}{\lambda} E(\tau_2^2)$$

Say that the effective ρ is 0.5, to be specific. The total arrival rate is then $\lambda = 6.17$ packets/sec, and using Eq. (2–65), the average waiting time for *either* type of packet is readily found to be 148 msec. The 48-bit control packets, requiring 5 msec for transmission, frequently may be trapped behind the much longer 100-msec data packets, and must wait, on the average, 148 msec for transmission! The obvious solution is to provide a higher priority to the control packets, enabling them to bypass the lower priority data packets on arrival and to move directly to the head of the queue.

Two types of priority are normally employed: nonpreemptive and preemptive. In the former case, higher-priority customers (packets in our example) move ahead of lower-priority ones in the queue but do not preempt lower-priority customers already in service. In the preemptive priority case, service on a lower-priority customer is interrupted and only continued after all arriving high-priority customers have been served. We focus here on the nonpreemptive priority case only.

Returning to the example just given, the case of short control packets and longer data packets, we shall show that providing nonpreemptive priority to the control packets essentially halves their wait time to 74.5 msec, on the average, while increasing the wait time of the data by an inconsequential amount. If this reduction is not sufficient in a real situation, preemptive priority may have to be employed. (There is a price paid, however: Customers whose service is interrupted must be so tagged. This entails added processing time and overhead, which may reduce some of the benefits theoretically obtained through preemption.)

To be more general, say there are r classes of customers to be served at a queue. Their respective arrival rates are $\lambda_1, \lambda_2, \ldots, \lambda_r$, each representing

a Poisson stream. The average service time is $1/\mu_k$ for the kth class, $k = 1, 2, \ldots, r$. The highest-priority class is taken to be 1, the lowest r, in descending order as labeled. We now show how one computes the average waiting time for any class, assuming nonpreemptive service. Consider class p $(1 \leq p \leq r)$ in particular. Let a typical customer of this class arrive at an arbitrary time t_0. Its random waiting time W_p (Fig. 2-30), measured from its arrival time until it enters service, is due to contributions from three sources. It must wait a random amount of time T_0 until the customer currently in service completes service. It must wait a random number T_k units of time until all customers of priority k lower than or equal to p, already enqueued at the arrival time t_0, complete service. Finally, it must wait a random time T'_k to service customers of each class k of priority lower than p arriving during the wait time W_p.

Putting these observations together, we write

$$W_p = T_0 + \sum_{k=1}^{p} T_k + \sum_{k=1}^{p-1} T'_k \tag{2-77}$$

Taking expectations term by term, the average waiting time $E(W_p)$ of priority p is obviously given by

$$E(W_p) = E(T_0) + \sum_{k=1}^{p} E(T_k) + \sum_{k=1}^{p-1} E(T'_k) \tag{2-78}$$

To find each of the three average times in Eq. (2-78), note first that $E(T_k)$ is due to an average number $E(m_k)$ customers of category k waiting in the system. Each requires $1/\mu_k$ units of service on the average, so that we immediately have

$$E(T_k) = E(m_k)/\mu_k \tag{2-79}$$

But by Little's formula, we also have $E(m_k)$ related to the average wait time $E(W_k)$. Specifically,

$$E(m_k) = \lambda_k E(W_k) \tag{2-80}$$

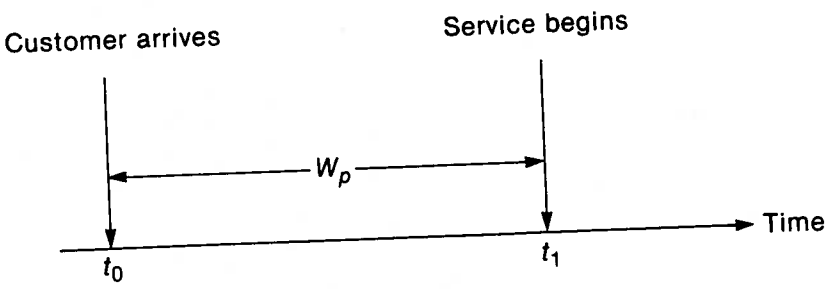

Figure 2-30 Waiting time, queueing system

Combining Eqs. (2–80) and (2–79), we immediately have

$$E(T_k) = \rho_k E(W_k) \qquad \rho_k \equiv \lambda_k/\mu_k \tag{2–81}$$

Now consider the term $E(T'_k)$ in Eq. (2–78). This is due, on the average, to $E(m'_k)$ customers of class k arriving during the interval $E(W_p)$. Since the arrival rate is λ_k and each customer again requires, on the average, $1/\mu_k$ units of service, we immediately have

$$E(T'_k) = \lambda_k E(W_p)/\mu_k = \rho_k E(W_p) \tag{2–82}$$

Consider the remaining term $E(T_0)$ in Eq. (2–78). This is the residual service time of a customer in service. For the work-conserving nonpreemptive queueing system under discussion here (the server always serves a customer if one is waiting to be served), this is independent of queue discipline: It must be the same if customers of all k classes are served with the same priority, in their order of arrival. From Eq. (2–65), the average wait time of an M/G/1 queue, we find that

$$E(T_o) = \lambda E(\tau^2)/2 = \sum_{k=1}^{r} \lambda_k E(\tau_k^2)/2 \tag{2–83}$$

This generalizes the two-class example described earlier.

Using Eqs. (2–82) and (2–81) in Eq. (2–78) and solving for the wait time of each class recursively, starting with the highest priority class 1, one readily shows that

$$E(W_p) = E(T_0)/(1 - \sigma_p)(1 - \sigma_{p-1}) \tag{2–84}$$

with $\sigma_p \equiv \Sigma_{k=1}^{p} \rho_k$, $\rho_k = \lambda_k/\mu_k$, and $E(T_o)$ given by Eq. (2–83).

As an example, consider the two highest-priority classes. For these we have

$$E(W_1) = E(T_0)/(1 - \rho_1) \tag{2–85}$$

and

$$\begin{aligned} E(W_2) &= E(T_0)/(1 - \rho_1)(1 - \rho) \\ &= E(W_1)/(1 - \rho) \qquad \rho = \rho_1 + \rho_2 \end{aligned} \tag{2–86}$$

The highest priority, class 1 customers, queue up as in an M/G/1 system of a single class, seeing themselves only, except for the added residual service time $E(T_0)$ that is due to customers of all classes that might have been in service.

In the special case of two priorities only, the case considered in the example described at the beginning of this section, Eqs. (2–85) and (2–86) provide the average waiting times for the two classes, respectively. Recall that with $\rho = 0.5$, and for the numbers given earlier, the average wait time $E(W)$ without priority is 148 msec. From Eq. (2–85), for the same example, we get $E(W_1) = 74.5$ msec. $E(W_2)$, using Eq. (2–86), turns out to be 149 msec. The average wait time of the

higher-priority control packets has thus dropped to almost half of the original value with no priority used, while the lower-priority data packets have had their wait time increased by 1 msec in 148, an inconsequential change. This demonstrates the improvement possible through the use of priority queueing.

Although the effect on the lower priority packets in this example is negligible, in other situations it might be much more noticeable. The fact is that as some priority classes (the higher ones) improve their performance, others deteriorate. Interestingly, it is simple to show that a conservation law is at work here. In particular, from Eq. (2–84), it is simple to show that the weighted sum of wait times is always conserved. In particular, one finds that

$$\sum_{k=1}^{r} \rho_k E(W_k) = \rho E(W) \tag{2-87}$$

with $E(W)$ given by Eq. (2–65), the wait time of the FIFO M/G/1 queue. As some wait times decrease, then, others must increase in order to compensate. As a check, in the two-priority example discussed here we had, without priority, $\rho E(W) = 74$ msec. With priority, $\rho_1 E(W_1) + \rho_2 E(W_2) = 74$ msec as well. This conservation law is a special case of a more general conservation law for work-conserving queues first developed by Kleinrock [KLEI 1965]. (See also [KLEI 1976].)

Problems

2–1 Refer to Fig. 2–3 in the text. Calculate the probability of k independent events in the m intervals Δt units long, if the probability of one event in any interval is p, while the probability of no event is $q = 1 - p$. Show how one obtains the binomial distribution of Eq. (2–4).

2–2 In Problem 2–1 let $p = \lambda \Delta t$, λ a proportionality factor. This then relates the binomial distribution to the Poisson process. Let $\Delta t \to 0$, with $T = m\Delta t$ fixed. Show that in the limit one gets the Poisson distribution of Eq. (2–1). Show that the mean value $E(k)$ and the variance σ_k^2 are both equal to λT. What is the probability that no arrival occurs in the interval T? Sketch this as a function of T. Repeat for the probability that *at least* one arrival occurs in T.

[KLEI 1965] L. Kleinrock, "A Conservation Law for a Wide Class of Queueing Systems," *Naval Res. Logist. Quarterly*, vol. 12, 1965, 181–192.

[KLEI 1976] L. Kleinrock, *Queueing Systems, Volume II: Computer Applications*, John Wiley & Sons, New York, 1976.

2-3 Calculate and plot the Poisson distribution given by Eq. (2-1) for the three cases $\lambda T = 0.1, 1, 10$. In the third case try to carry the calculation and plot out to at least $k = 20$. (Stirling's approximation for the factorial may be useful here.) Does the distribution begin to crowd in and peak about $E(k)$ as predicted by the ratio $\sigma_k/E(k) = 1/\sqrt{\lambda T}$?

2-4 Carry out the details of the analysis leading to Eqs. (2-10) and (2-11), showing that sums of Poisson processes are Poisson as well.

2-5 Refer to the time-dependent equation (2-12a) governing the operation of the M/M/1 queue. Start at time $t = 0$ with the queue empty. (What are then the values $p_n(0)$?) Let $\lambda/\mu = 0.5$ for simplicity, take $\Delta t = 1$, and pick $\lambda\Delta t$ and $\mu\Delta t$ very small so that terms of $(\Delta t)^2$ and higher can be ignored. Write a program that calculates $p_n(t + \Delta t)$ recursively as t is incremented by Δt and show that $p_n(t)$ does settle down eventually to the steady-state set of probabilities $\{p_n\}$. Pick the maximum value of n to be 5. The set of steady-state probabilities obtained should then agree with Eq. (2-20). *Note:* Eq. (2-12a) must be modified slightly in calculating $p_0(t + \Delta t)$ and $p_5(t + \Delta t)$. You may want to set the problem up in matrix-vector form.

2-6 Derive Eq. (2-15), governing the steady-state (stationary) probabilities of state of the M/M/1 queue, in two ways:

1. from the initial generating equation (2-12)
2. from flow balance arguments involving transitions between states $n - 1$, n, and $n + 1$, as indicated in Fig. 2-11.

2-7 As a generalization of the M/M/1 queue analysis, consider a birth-death process with state-dependent arrivals λ_n and state-dependent departures μ_n. (See Figs. 2-24 and 2-25.) Show, by applying balance arguments, that the equation governing the stationary state probabilities is given by

$$(\lambda_n + \mu_n)p_n = \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}$$

(See Eq. (2-38).) Show that the solution to this equation is given by Eq. (2-40).

2-8 Consider the M/M/1 queue analysis. Show that the stationary state probability p_n is given by

$$p_n = \rho^n p_0 \qquad \rho \equiv \lambda/\mu$$

in two ways:

1. Show that this solution for p_n satisfies Eq. (2-15) governing the queue operation.
2. Show that the balance equation $\lambda p_n = \mu p_{n+1}$ or $p_{n+1} = \rho p_n$ satisfies Eq. (2-15). Then iterate n times.

Calculate p_0 for the finite M/M/1 queue and show that p_n is given by Eq. (2-20).

2-9 Show that the blocking probability P_B of the finite M/M/1 queue is given by $P_B = p_N$ by equating the net arrival rate $\lambda(1 - P_B)$ to the average departure rate $\mu(1 - p_0)$ and solving for P_B.

2–10 Consider a finite M/M/1 queue capable of accommodating N packets (customers). Calculate the values of N required for the following situations:

1. $\rho = 0.5$, $P_B = 10^{-3}$, 10^{-6}
2. $\rho = 0.8$, $P_B = 10^{-3}$, 10^{-6}

Compare the results obtained.

2–11 The probability p_n that an infinite M/M/1 queue is in state n is given by $p_n = (1 - \rho)\rho^n \quad \rho = \lambda/\mu$.

a. Show that the average queue occupancy is given by

$$E(n) = \sum_n np_n = \rho/(1 - \rho)$$

b. Plot p_n as a function of n for $\rho = 0.8$.

c. Plot $E(n)$ versus ρ and compare with Fig. 2–17.

2–12 The average buffer occupancy of a statistical multiplexer (or data concentrator) is to be calculated for a number of cases. (In such a device the input packets from terminals connected to it are merged in order of arrival in a buffer and are then read out first come–first served over an outgoing transmission link.) An infinite buffer M/M/1 model is to be used to represent the concentrator.

1. Ten terminals are connected to the statistical multiplexer. Each generates, on the average, one 960-bit packet, assumed to be distributed exponentially, every 8 sec. A 2400-bits/sec outgoing line is used.
2. Repeat if each terminal now generates a packet every 5 sec, on the average.
3. Repeat 1. above if 16 terminals are connected.
4. Forty terminals are now connected and a 9600-bits/sec output line is used. Repeat 1. and 2. in this case. Now increase the average packet length to 1600 bits. What is the average buffer occupancy if a packet is generated every 8 sec at each terminal? What would happen if each terminal were allowed to increase its packet generation rate to 1 per 5 sec, on the average? (*Hint:* It might now be appropriate to use a *finite* M/M/1 model with your own choice of buffer size.)

2–13 Consider the finite M/M/1 queue holding at most N packets.

a. Show that the blocking probability is $P_B = 1/(N + 1)$ at $\rho = 1$.

b. Plot $P_B = p_N$ for all values of ρ, $0 \leq \rho < \infty$ for $N = 4$ and $N = 19$.

c. The throughput is defined to be $\gamma = \lambda(1 - P_B)$. Sketch γ/μ (the normalized throughput) as a function of $\rho = \lambda/\mu$ (normalized load) for $0 \leq \rho < \infty$, $N = 4$, and $N = 19$. Compare.

2–14 Refer to Problem 2–12. Find the mean delay $E(T)$ and the average wait time $E(W)$ in each case.

2–15 Use Fig. 2–23 to prove that Little's formula applies in the LCFS service discipline.

2–16 Draw your own arrival-departure diagrams for an arbitrary queueing system, as

in Figs. 2–22 and 2–23, comparing the FIFO and LCFS service disciplines. Carry out your own proof of Little's formula.

2–17 Consider the M/M/2 queue discussed in the text. Derive Eq. (2–43), the expression for the probability of state occupancy and Eq. (2–44), the equation for the average queue occupancy. Plot $\mu E(T)$ (normalized time delay) versus λ, the average arrival rate (load on the system) for the M/M/2 queue, and compare with two single-server cases: an M/M/1 queue with service rate μ and an M/M/1 queue with service rate 2μ. Check Fig. 2–27.

2–18 Refer to the multiple (or *ample*) server and queue with discouragement examples discussed in the text. Show that the state probability distribution and the average queue occupancy are given, in both examples, by Eq. (2–47) with Eq. (2–48), and Eq. (2–49), respectively. However, the average time delay and throughput are different in the two cases. Calculate these two quantities in both cases and compare.

2–19 Consider a queue with a general state-dependent departure (service) process μ_n.

a. Explain why the average throughput is given by $\gamma = \sum_{n=1}^{N} \mu_n p_n$.

b. Take the special case of the M/M/2 queue, $\mu_n = \mu$, $n = 1$; $\mu_n = 2\mu$, $n \geq 2$. Show that $\gamma = \mu p_1 + 2\mu(1 - p_0 - p_1)$. Show that this is just $\gamma = \lambda$, if p_1 and p_0 are explicitly calculated using Eq. (2–41).

2–20 A queueing system holds N packets, including the one(s) in service. The service rate is state dependent, with $\mu_n = n\mu$, $1 \leq n \leq N$. Arrivals are Poisson, with average rate λ.

a. Show that the probability the system is in state n is the Erlang distribution of Eq. (2–54).

b. Show that the average number in the system is given by Eq. (2–56).

c. Show that the average *throughput* is $\gamma = \mu E(n)$, in two ways:

1. Use $\gamma = \sum_{n=0}^{N} \mu_n p_n$
2. $\gamma = \lambda(1 - P_B)$, P_B the blocking probability.

d. Little's theorem says that $E(T) = E(n)/\gamma = 1/\mu$ here. Explain this result (i.e., there is *no* waiting time).

2–21 A queueing system has two outgoing lines, used randomly by packets requiring service. Each transmits at a rate of μ packets/sec. When both lines are transmitting (serving) packets, packets are blocked from entering—i.e., there is no buffering in this system. Packets are exponentially distributed in length; arrivals are Poisson, with average rate λ. $\rho = \lambda/\mu = 1$.

a. Find the blocking probability, P_B, of this system.

b. Find the average number, $E(n)$, in the system.

c. Find the normalized throughput γ/μ, with γ the average throughput, in packets/sec.

d. Find the average delay $E(T)$ through the system, in units of $1/\mu$. (Alternatively, find $E(T)/1/\mu$.)

2–22 A data concentrator has 40 terminals connected to it. Each terminal inputs

packets with an average length of 680 bits. Forty bits of control information are added to each packet before transmission over an outgoing link with capacity $C = 7200$ bps.

Twenty of the terminals input 1 packet/10 sec each, on the average.
Ten of the terminals input 1 packet/5 sec each, on the average.
Ten of the terminals input 1 packet/2.5 sec each, on the average.
The input statistics are Poisson.

a. The data units transmitted (called frames) are exponentially distributed in length. Find (1) the average wait time on queue, *not including* service time and (2) the average number of packets in the concentrator, *including* the one in service.

b. Repeat if the packets are all of constant length.

c. Repeat if the second moment of the frame length is $E(\tau^2) = 3(1/\mu)^2$; $1/\mu$ is the average frame length.

2–23 Refer to the derivation of the Pollaczek-Khinchine formulas in the text.

a. Derive Eq. (2–69), the general expression for the average number of customers in the queue.

b. For the case of Poisson arrivals, calculate $E(\nu)$ and σ_ν^2, following the procedure in the text, and show that Eqs. (2–74) and (2–76) result.

2–24 Two types of packets are transmitted over a data network. Type 1, control packets, are all 48 bits long; type 2, data packets, are 960 bits long on the average. The transmission links all have a capacity of 9600 bps. The data packets have a variance $\sigma_2^2 = 2(1/\mu_2)^2$, with $1/\mu_2$ the average packet length in seconds. The type 1 control packets constitute 20 percent of the total traffic. The overall traffic utilization over a transmission link is $\rho = 0.5$.

a. FIFO (nonpriority service) is used. Show that the average waiting time for either type of packet is $E(W) = 148$ msec.

b. Nonpreemptive priority is given to the control packets (type 1). Show that the wait time of these packets is reduced to $E(W_1) = 74.5$ msec, whereas the wait time of the data packets (type 2) is increased slightly to $E(W_2) = 149$ msec.

2–25 Show that the average wait time of class p in a nonpreemptive priority system is given by Eq. (2–84).